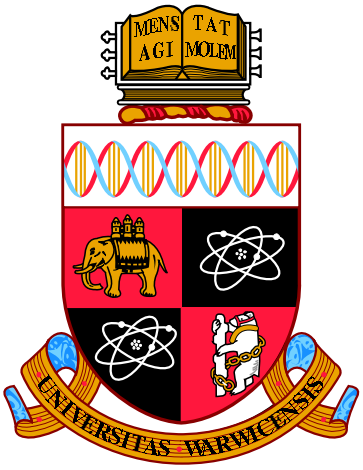

On Sample Selection Models and Skew Distributions

by

Emmanuel Olusegun Ogundimu

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

February 2013

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

Let me begin by expressing my deepest gratitude to my supervisor, Prof. Jane Luise Hutton for her support and encouragement which enabled me to complete this thesis. She believes in me as a mother would do with her child. Thanks for introducing me to skew distributions and pointing me in the right direction to better my future career in statistics. The technical support, and time given to me during the preparation of the thesis is invaluable. The four Xmas celebrated during the period of my studentship were celebrated home away from home at her place. Thanks for the tasty dishes. May you live long to eat the fruits of your labor. I sincerely look forward to future collaboration with you because no one ever forgets a good teacher. Thank you again and again!

I am very grateful to my examiners, Dr. Ewart Shaw and Prof. David Firth. The feedback from Dr. Shaw on my second year report improved the final draft of this thesis. Prof. Firth's critical view on ceiling and floor effects of bounded scores on likelihood-based inference motivated the use of truncated distributions in the thesis. A big thank you to Prof John Copas for helpful insight that improved this work. I wish him all the very best in his retirement. Thanks to Prof. Sallie Lamb and the MINT trial team for the permission to use the Neck disability index data. My PhD research was funded by the Engineering and Physical Sciences Research Council grant, for the Centre for Research in Statistical Methodology. I am grateful to the Department of Statistics for this grant.

I received a very warm reception and hospitality during my three months placement at Novartis Pharma, Switzerland. Many thanks to Mouna Akacha for all the stimulating discussions on models for recurrent event data. I appreciate Prof.

Frank Bretz for the opportunity to be part of 'the Novartis dream'. May God bless you.

Many people contributed to my academic career till date. I thank all my teachers. In particular, Prof. Geert Molenberghs who has been of tremendous assistance. He is ever ready to give me his shoulder to stand on and see farther. I appreciate my Professor and mentor, Prof. Adewale R. Solarin for his support. I appreciate my former Heads of Department, Prof. Kaku Sagary Nokoe and Prof. James Adedayo Oguntuase for their support and love. May the Lord reward you.

My appreciation is definitely incomplete without mentioning friends and family. I appreciate my Mum that taught me that hard work and perseverance pay. She is fond of our local version of the French proverb, 'One may go a long way after one is tired'. Out of sight is indeed not always out of mind. A big thanks to my uncle, Johnson Sunday Olawunmi, for his incessant calls to check how I am fairing during the last three and a half years. It is of course true that behind every successful man, there is a woman. Thanks to my angel Oluwakemi Racheal Adeboye for her patience and unalloyed support throughout the period of my PhD research. You are an angel indeed! Dr. Peter Kimani was my first Warwick friend and he has been of enormous support. I am very grateful Peter, God bless you. I had interesting discussions with my colleague, Javier Rubio on skew distributions. Thank you Javier. Alex Thiery was and is still a good friend and colleague. God bless you.

Above all, I appreciate God who was my help in ages past, my hope for years to come, my shelter from the stormy blast, and my eternal home.

Declarations

I declare that the work in this thesis is my own, and has not been submitted elsewhere for examination. The materials that are not my original ideas have been acknowledged by referencing. The work in Chapter 7 was done jointly with Mouna Akacha during my internship at Novartis Pharma, Switzerland.

Abstract

This thesis is concerned with methods for dealing with missing data in non-random samples and recurrent events data.

The first part of this thesis is motivated by scores arising from questionnaires which often follow asymmetric distributions, on a fixed range. This can be due to scores clustering at one end of the scale or selective reporting. Sometimes, the scores are further subjected to sample selection resulting in partial observability. Thus, methods based on complete cases for skew data are inadequate for the analysis of such data and a general sample selection model is required. Heckman proposed a full maximum likelihood estimation method under the normality assumption for sample selection problems, and parametric and non-parametric extensions have been proposed.

A general selection distribution for a vector $\mathbf{Y} \in \mathbb{R}^p$ has a PDF f_Y given by

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}^\star}(\mathbf{y}) \frac{P(\mathbf{S}^\star \in \mathbf{C} \,|\, \mathbf{Y}^\star = y)}{P(\mathbf{S}^\star \in \mathbf{C})},$$

where $\mathbf{S}^\star \in \mathbb{R}^q$ and $\mathbf{Y}^\star \in \mathbb{R}^p$ are two random vectors, and $\mathbf{C}$ is a measurable subset of $\mathbb{R}^q$. We use this generalization to develop a sample selection model with underlying skew-normal distribution. A link is established between the continuous component of our model log-likelihood function and an extended version of a generalized skew-normal distribution. This link is used to derive the expected value of the model, which extends Heckman's two-step method. The general selection distribution is also used to establish the closed skew-normal distribution as the continuous component of the usual multilevel sample selection models. Finite sample performances of the maximum likelihood estimator of the models are studied via Monte Carlo simulation. The model parameters are more precisely estimated under the new models, even in the presence of moderate to extreme skewness, than the Heckman selection models. Application to data from a study of neck injuries where the responses are substantially skew successfully discriminates between selection and inherent skewness, and the multilevel model is used to analyze jointly unit and item non-response. We also discuss computational and identification issues, and provide an extension of the model using copula-based sample selection models with truncated marginals.

The second part of this thesis is motivated by studies that seek to analyze processes that generate events repeatedly over time. We consider the number of events per subject within a specified study period as the primary outcome of interest. One considerable challenge in the analysis of this type of data is the large proportion of patients that might discontinue before the end of the study, leading to partially observed data. Sophisticated sensitivity analyses tools are therefore necessary for the analysis of such data.

We propose the use of two frequentist based imputation methods for dealing with missing data in recurrent event data framework. The recurrent events are modeled as over-dispersed Poisson data, with constant rate function. Different assumptions about future behavior of dropouts depending on reasons for dropout and treatment received are made and evaluated in a simulation study. We illustrate our approach with a clinical trial in patients who suffer from bladder cancer.

Abbreviations

base	Baseline measurement for the NDI scores
CDF	Cumulative Distribution Function
CSN	Closed Skew-Normal
DL	Direct Likelihood
EGSN	Extended (two-parameter) Generalized Skew-Normal
ESN	Extended Skew-Normal
int	Intercept
$l'HR$	l'Hospital's Rule
Loglik	Log-likelihood value
LRT	Likelihood Ratio Test
MAR	Missing At Random
MCAR	Missing Completely At Random
mgf	Moment Generating Function
MI	Multiple Imputation
MLEs	Maximum Likelihood Estimates
MNAR	Missing Not At Random
MSN	Multivariate Skew-Normal (Azzalini's Skew-normal distribution)
NDI	Neck Disability Index
Num	Number of tumor (Bladder cancer data)
PDF	Probability Density Function
Physio	Physiotherapy treatment
pMI	Placebo Multiple Imputation
Prev	Previous Measurement (Measurements at Month 4)

Q-Q plot	Quantile-Quantile plot
S.E	Standard Error
SHASH	Asymmetric subfamily of Sinh-Arcsinh distribution
Size	Size of tumor (Bladder cancer data)
SN	Skew-Normal (Azzalini's univariate Skew-normal distribution)
SNM	Selection Normal Model
SSNM	Selection Skew-Normal Model
SUN	Unified Skew-Normal
SUT	Unified Skew-t
TS	Heckman Two-step Method
TSN	Truncated Skew-Normal
WAD	Whiplash Associated Disorder

Chapter 1

Introduction

This thesis discusses issues arising with missing data, in two parts. The first part is devoted to the unification of missing data problems into a distributional framework, while the second part considers a distinct, but related, concept of dealing with missing data in a recurrent events data framework.

The first part of the thesis is motivated by a study where pain related activity restriction is measured repeatedly over time using the *neck disability index* (NDI) questionnaire (Vernon and Mior, 1991). In this type of study, the patient's perception of his or her well-being is usually the most important outcome of interest. These are broadly termed quality of life (QoL) outcomes. Scores arising from instruments designed to assess QoL (e.g. screening questionnaires) often follow asymmetric distributions due to skewness inherent in Likert-scale type instruments. Indeed, skewness related studies are not uncommon in psychology literature. In addition, the realized samples from the underlying discrete process are further subjected to selective reporting and missing data, with the scores reflecting a selected population. Consequently, there is need for a general model for sample selection with inherent skewness.

The two most common deviations from normality are heavier tails and skewness. In dealing with heavier tails in sample selection, Marchenko and Genton (2012) derived a model using links between hidden truncation and sample selection but with an underlying bivariate-t error distribution. They noted that a more appealing flexible parametric model is needed to be considered that can accommodate heavy tails and skewness. A skew-normal distribution (Azzalini, 1985) could be a good candidate to accommodate skewness.

An additional, commonly observed complication in the analysis of QoL study is that they are usually planned as longitudinal studies. Sometimes, the treatment

effects at a measurement occasion may be desirable and a cross-sectional view of the data will make two missing data type inevitable- unit and item non-response. Unit non-response occurs when the whole questionnaire is missing for a patient and item non-response occurs where a response has not been provided for a question. The traditional practice is to use weighting adjustment for unit non-response and imputation methods for item non-response. Weighting adjustment means weights are assigned to sample respondents in order to compensate for their systematic differences relative to non-respondents, whereas imputation involves filling in missing values (singly or multiply) to produce complete data set.

Although these methods have reached a high level of sophistication, they normally assume that the missing data mechanism is missing at random (MAR), an assumption that cannot be verified using the observed data alone. Apart from this, patients may refuse to answer sensitive questions (e.g. underlying health issues or drug addiction) on a questionnaire for reasons related to the underlying true values for those questions. In multivariate settings with arbitrary patterns of non-response, imputation, and hence the MAR assumption, is convenient computationally, but it is often implausible (Robins and Gill, 1997). In this setting, MAR means that a patient's probabilities of responding to items may depend only on his or her own set of observed items, which is an unrealistic assumption. Specifically, the use of mean imputation is justifiable if items within the scale are strongly correlated with each other but correlation with external factors is low relative to within-scale correlations. This cannot be readily established in practice. Thus, when we suspect that non-response may depend on missing values, then a proper analysis will be to model jointly the population of complete data and the non-response process. Sample selection models are therefore viable tool.

A selection model was introduced by Heckman (1976). He proposed a full maximum likelihood estimation under the assumption of normality. His method was criticized on the ground of its sensitivity to normality assumption prompting him to develop the two-step estimator (Heckman, 1979). Sample selection models, also referred to as models with incidental (hidden) truncation, arise in practice as a result of the partial observability of the outcome of interest in a study. The data are missing not at random (MNAR) because the observed data do not represent a random sample from the population, even after controlling for covariates. Although the model has its origin from the field of Economics, it has been applied extensively in other social sciences, and in medicine. A prominent application to treatment allocation for patients and links with the skew-normal distribution was discussed by Copas and Li (1997).

There are situations where a variable is skewed and yet the residuals are approximately normal when the skewed variable is conditioned on other variables. This however, is not the case with bounded scores since the data exhibits ceiling and floor effects and the skewness could be natural consequences of this. The classical approach is to transform the data to near normality so that a linear regression model can be used. This may not remove the non-linear dependence of the transformed scores on covariates because of the bounds (see Hutton and Stanghellini (2011)). In fact, if such transformations exist, they are not always appropriate in modeling data resulting from selectively reported samples because interest is in making inference in the unselected population. There is additional disadvantage of not working on the original scale familiar to the health care professionals.

In view of these limitations, we propose extensions of Heckman (1976) and Heckman (1979) models by adding two additional features in a parametric framework. First, a skew-normal error distribution is used as an underlying error distribution. This model allows us to establish a link between the continuous component of our model log-likelihood function and an extended version of a generalized skew-normal distribution (Jamalizadeh et al., 2008). Sensitivity analysis for the assumption of selection is readily carried out using the profile likelihood in a manner similar to the Copas and Li (1997) approach. In addition, the link is used to derive the expected value of the model, which extends Heckmans two-step method. Secondly, sample selection model is unified into a distributional framework. This allows for straightforward extensions of Heckman's models into multilevel and longitudinal framework. In particular, the model is used to analyze jointly a data set with unit and item non-response. Sample selection models using Gaussian copula are also investigated.

The second part of this thesis is motivated by a study that compares an active treatment with a placebo in a recurrent event data framework, subject to informative dropout. The aim is to provide a tool for sensitivity analysis in such studies. Recurrent event data arise in practice when a subject experiences the same type of event repeatedly over time. Unlike in a classical survival study where patients can experience at most a single event, patients can experience multiple events in recurrent event data framework. For example, in clinical research, repeated seizures in epileptic patients, flares in gout studies or repeated asthma attacks can be classified as recurrent events.

A point process formulation is commonly used to describe and analyse recurrent event data and the two most commonly used approaches are the event counts or gap/ waiting times between successive events (Cook and Lawless, 2007). Models

based on event counts are used to describe situations where events occur randomly in such a way that the numbers of events in non-overlapping time intervals are statistically independent. These models are often used for frequently occurring events in a subject. On the other hand, the gap time approaches are often used when events are relatively infrequent. This method is ideal for situations where prediction of time to next event is of interest, and is very common in studies that investigate system failures. Our focus in this part of the thesis will be on event counts and the traditional framework for its analysis, the Poisson process.

Recurrent event data analysis takes the whole evolution of the recurrent events into account. There are potential problems in the presence of dropout. First, if we assume an *intention to treat analysis (ITT, i.e. patients data are analyzed in the treatments groups they are randomized to and not on the treatments they eventually received)* we need to take into account the follow-up time. This is because the number of events may be the same for two patients but the number of counts per unit time, (i.e. number of count/follow-up time) may differ substantially. For example, a patient who drops out, say, after the second event, due to toxicity has event count of two. On the other hand, there might be less dropout in the placebo group with high number of events. Thus, the treatment might appear to be effective when in fact the latent reason is the high dropout rate in the treated group. Of course, the dropout time can be adjusted for in the model and this will give valid analysis if the missingness process is unrelated with the outcome process. This does not give sufficient flexibility to examine other types of missing data mechanism that can also bias the treatment comparison.

Consequently, we examine in a simulation study how data analyses results can depend on assumptions of MAR and MNAR, and the imputation methods used to impute the missing data. The flexibility and transparency of multiple imputation makes it attractive for this work. In addition, multiple imputation separates the solution of the missing data problem from the solution of the complete data problem. The missing data problem is first solved before solving the complete data problem. The fact that these two phases can be separated gives a better insight into the scientific problems we study in this part of the thesis. We also investigate the importance of varying event generation process (see (Metcalfe and Thompson, 2006; Jahn-Eimermacher, 2008)) and the impact of the imputation methods used.

1.1 Overview of Thesis

The thesis has eight chapters and is organized in two parts. The first part is motivated by the MINT trial (Managing Injuries of the Neck Trial) which uses the NDI scores, and the second is motivated by a publicly available bladder cancer data set. In the introductory part, we pointed out that selectively reported outcomes often leads to skewness. This selectivity may result not only from decisions on sampling design but also from self-selection. An overview of relevant literature on skew-normal distribution is provided in chapter 2. Given that the univariate and multivariate normal distributions are well known, we will assume that the underlying process follows normal laws. Since selection under the normal process leads to the familiar Azzalini (1985) (or its extension) skew-normal distribution, this allows us to describe their connections with missing data. Exploratory analysis of the data set used in this part of the thesis and the concept of missing data concludes this chapter.

Methods that ignore the missing data process are discussed in chapter 3. In particular, we introduce a new class of skew-normal distribution which we referred to as an *extended two-parameter generalized skew-normal distribution*. The implication of using skew-normal distributions to model data arising from sample selection is evaluated in a simulation study, and data example concludes this chapter.

In chapter 4, we develop a sample selection model with underlying skew-normal distribution which we referred to as selection skew-normal model (SSNM). Its moment estimator was derived using the link between skew models arising from selection and hidden truncation formulation of skew models. The moment estimator is shown to extend Heckman two-step method. A simulation study is used to demonstrate the superiority of the SSNM model over the conventional sample selection model and data application is considered. We conclude this chapter by proposing a multivariate extension of this model in a straightforward way.

In chapter 5, we propose a unified approach for multilevel sample selection models in a parametric framework by treating the outcome variable as the non-truncated marginal of a truncated multivariate normal distribution. The resulting density for the outcome is the continuous component of the sample selection density, and has links with the closed skew-normal distribution. The closed skew-normal distribution provides a framework which simplifies the derivation of the conditional expectation and variance of the observed data. We use this to generalize the Heckman's two-step method to a multilevel sample selection model. This model is used to analyze jointly unit and item non-response in the NDI scores.

A major draw-back of the model proposed in chapter 4 is that a solution to the score equations always exists associated with the skewness parameter equals to zero. This feature is inherited from the underlying Azzalini (1985) skew-normal distribution used. To circumvent this problem, we propose in chapter 6, the use of Gaussian copula in a sample selection framework with the Jones and Pewsey (2009) sinh-arcsinh distribution as marginals. We examine the power of Wald test and LRT for the hypothesis of symmetry. We conclude the chapter with the examination of the impact of boundedness in the NDI scores on inherent skewness in the data using sample selection models with truncated marginal distributions for the outcomes.

The second part of this thesis focus on imputation of missing data in recurrent event data study. We propose a method for artificially creating the missing recurrent event sequence for the data under the assumption that patients get no benefit if they stop taking the active treatment. This method of imputation is referred to as placebo multiple imputation (pMI). The MAR assumption implies that the future statistical behavior of the observations from a subject, conditional on the history, is the same whether the subject drops out (deviates) or not in the future. Based on this, we propose sensitivity analysis tools in a simulation study by imputing missing data for patients in the active treatment with higher event rate than the one determined by the MAR assumption. In chapter 7, we review models for recurrent event data, and two frequentist based imputation methods are evaluated. To make the method readily available to applied statisticians, we give an easy to follow algorithm to execute the imputation model. A scenario evaluation study to compare the performances of the methods proposed in this part is also studied. A data example completes this chapter.

In chapter 8, an overall conclusion of this thesis and direction for future research is presented.

Part I

On Sample Selection Models and Skew-Normal Distributions

Chapter 2

Literature Review

There is an enormous body of literature that address skew distributions and sample selection separately and jointly. Arguably, most of the works are very general and well grounded mathematically. However, these works have been applied sparingly in modeling real life data. We present the Azzalini (1985) skew-normal distribution and its links with sample selection problems. Other methods for the construction of skew distributions are discussed. In addition, we introduce the data set that is used in this part of the thesis. Data exploration which motivated the models proposed in the thesis is also evaluated. Concepts of missing data conclude this chapter.

2.1 Skew-Normal Distribution

The skew-normal distributions are extensions of the normal distribution which admit skewness whilst retaining most of the interesting properties of the normal distribution. Their popularity, since the Azzalini (1985) paper, has led to intense development of this class. The developments are so numerous that it is confusing to applied statisticians which class of skew-normal model is most appropriate for data analysis. The relationship between these models are discussed below.

2.1.1 Univariate Skew-normal distribution

A random variable (r.v) Z is said to have a skew symmetric distribution generated by g and π, if its probability density function (PDF) is

$$f_Z(z) = 2g(z)\pi(z), \quad z \in \mathbb{R}, \tag{2.1}$$

where g is a PDF symmetric about 0 and π is a Lebesgue measurable function satisfying $0 \leq \pi(z) \leq 1$ and $\pi(z) + \pi(-z) = 1$, almost everywhere on $\mathbb{R}$. The function π is called a skewing function.

Skew-symmetric distributions have been investigated by many authors. For various π, Nadarajah and Kotz (2003) and Arellano-Valle et al. (2004) studied the properties of skew-symmetric distributions with $g = \phi$, the standard normal density. The cases in which $\pi(z) \equiv \Psi(\lambda z)$, λ, $z \in \mathbb{R}$ where Ψ is a CDF with Ψ' symmetric about 0, and g is any of the following PDFs: normal, Student's t, Cauchy, Laplace, logistic, and uniform has also been investigated (see Gupta et al. (2002)).

The theory of skew-symmetric distributions begins with the Azzalini (1985) paper where $g(z) = \phi(z)$ is combined with the skewing function $\pi(z) = \Phi(\lambda z)$, where Φ denotes the standard normal CDF.

Definition 1. *Let Z be a continuous random variable. Let ϕ and Φ denote the standard normal density and corresponding distribution function respectively. Then Z is said to have a skew-normal distribution with parameter $\lambda \in \mathbb{R}$ if the density of Z is*

$$f(z;\lambda) = 2\phi(z)\Phi(\lambda z), \quad z \in \mathbb{R} \tag{2.2}$$

and we write $Z \sim SN(\lambda)$.

The component λ is called the shape parameter because it regulates the shape of the density function. When $\lambda = 0$, the density is the standard normal. Figure 2.1 shows the densities corresponding to 4 different positive skewness. It can be seen that the model converges to half-normal distribution very fast as λ increases, even for values of λ as small as 5 or 10. In practice, to fit data, we work with an affine transformation $Y = \mu + \sigma Z$, $\mu \in \mathbb{R}$ and $\sigma > 0$. The density of Y is then written as

$$f(y;\mu,\sigma,\lambda) = \frac{2}{\sigma}\phi\Big(\frac{y-\mu}{\sigma}\Big)\Phi\Big(\lambda\frac{y-\mu}{\sigma}\Big), \tag{2.3}$$

and we write $Y \sim SN(\mu,\sigma,\lambda)$. A convolution type stochastic representation of (2.2) in terms of a normal and a half normal was given by Henze (1986). If Y_0 and Y_1 are independent $N(0,1)$ random variables and $\delta \in [-1,1]$, then

$$Z = \delta|Y_0| + \sqrt{1-\delta^2}Y_1,$$

is $SN(\lambda)$, where $\lambda = \delta/\sqrt{1-\delta^2}$.

Some important properties of the density include:

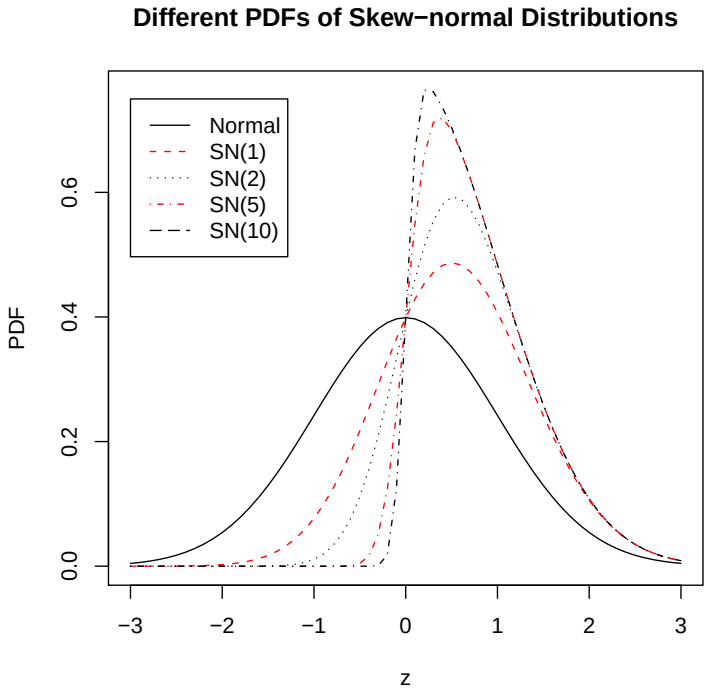

Figure 2.1: Comparison of Skew-normal densities

- $E(Z) = \lambda\sqrt{2/\pi}$
- Var(Y)= $1 - \frac{2}{\pi\lambda^2}$
- Skewness index $\gamma = \Big(2/\pi\Big)^{3/2}\Big(2-\pi/2\Big)\text{sign}(\lambda)\lambda^2\Big/\Big(1-2\lambda^2/\pi\Big)^{3/2} \in$ [-0.995,0.995].

The CDF of (2.2) is

$$2\int_{-\infty}^{z}\int_{-\infty}^{\lambda s}\phi(s)\phi(t)\mathrm{d}t\mathrm{d}s = 2\Phi_2\Big(z, 0; -\lambda\Big/\sqrt{1+\lambda^2}\Big),$$

where Φ_2 is the CDF of a standard bivariate normal distribution.

The skew-normal distribution and its multivariate counterparts suffer from two inferential drawbacks. When the skewness parameter equals zero, the profile likelihood for skewness admits stationary points for any sample of any size, and the Fisher information matrix is singular. These problems have not limited the usefulness of the distribution in practice (see Pewsey (2000), Ley and Paindaveine (2010) and Hallin and Ley (2012)).

2.1.2 Multivariate Skew-normal distribution (MSN)

The multivariate skew-normal distribution, like its univariate counterpart, has some properties similar to the normal distribution and includes the normal distribution as a special case.

Definition 2. *A random vector* $\mathbf{Z}= (Z_1, \ldots, Z_p)'$ *is a p-dimensional skew-normal, denoted* $\mathbf{Z} \sim SN_p(\bar{\Omega}, \boldsymbol{\lambda})$*, if it is continuous with PDF*

$$f(\mathbf{z}) = 2\phi_p(\mathbf{z}; \bar{\Omega})\Phi(\boldsymbol{\lambda}'\mathbf{z}), \quad \mathbf{z} \in \mathbb{R}^p \tag{2.4}$$

where $\phi_p(\mathbf{z}; \bar{\Omega})$ *denotes the PDF of the p-dimensional multivariate normal distribution with standardized marginals and correlation matrix* $\bar{\Omega}$*.*

If $p = 2$, the PDF given in (2.4) becomes

$$f(z_1, z_2) = 2\phi_2(z_1, z_2; \omega)\Phi(\lambda_1 z_1 + \lambda_2 z_2), \tag{2.5}$$

where ω is the off-diagonal element of $\bar{\Omega}$. As in the univariate case, when a location-scale transformation of the type $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{SZ}$ is applied, we have the PDF of Y as

$$f(\mathbf{y}) = 2\phi_p(\mathbf{y}; \boldsymbol{\mu}, \Omega)\Phi(\boldsymbol{\lambda}'\mathbf{S}^{-1}(\mathbf{y} - \boldsymbol{\mu})),$$

where $\Omega = \mathbf{S}\bar{\Omega}\mathbf{S}$, and we write $\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$, $\mathbf{S} = diag(\sigma_1, \ldots, \sigma_p)$. Like the univariate SN, density 2.4 has some attractive properties:

- If $\boldsymbol{\lambda} = \mathbf{0}$, then the model reduces to standard multivariate normal.
- If $\mathbf{Y} \sim N_p(\mathbf{0}, \bar{\Omega})$ and $\mathbf{Z} \sim SN_p(\bar{\Omega}, \boldsymbol{\lambda})$, then $\mathbf{Y}'\bar{\Omega}^{-1}\mathbf{Y}$ and $\mathbf{Z}'\bar{\Omega}^{-1}\mathbf{Z}$ have the same distribution i.e. χ^2_p
- If $\mathbf{Z} \sim SN_p(\bar{\Omega}, \boldsymbol{\lambda})$ and B is a symmetric positive semi-definite $p \times p$ matrix of rank k such that $B\bar{\Omega}B = B$, then $\mathbf{Z}'B\mathbf{Z} \sim \chi^2_k$.

Details on how to generate MSN distribution including multivariate generalization of Henze (1986) can be found in Genton (2004).

The contours of the bivariate skew-normal density are not elliptical (see Figure 2.2). This implies that the correlation coefficient is not a good measure of association between the two bivariate variables. The implication of this will be discussed in chapter 4. Although the distributions have properties similar to the normal distribution, they lack the important property of closure under conditioning as the following Theorem shows.

Theorem 1. *Let* $(Z_1, Z_2)' \sim SN_2$. *The conditional density* $f(Z_2|Z_1 = z_1)$ *is*

$$\frac{\phi_c(z_2|z_1;\omega)\Phi(\lambda_1 z_1 + \lambda_2 z_2)}{\Phi(\lambda_1 z_1)}, \tag{2.6}$$

where $\phi_c(z_2|z_1;\omega)$ *denotes the conditional density associated with a bivariate normal variable with standardized marginals and correlation* ω.

Equation (2.6) belongs to the extended skew-normal (ESN) family (Azzalini and Dalla Valle, 1996; Capitanio et al., 2003).

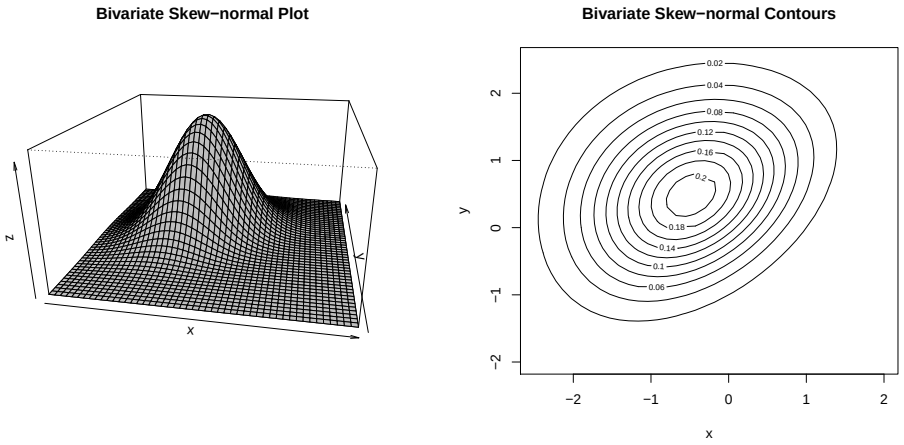

Figure 2.2: Contour plot and 3-d plot of a bivariate $SN_2(\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda})$ with $\boldsymbol{\mu} = (-0.1, 0.1)$, $\Omega = \text{diag}(1,1)$ and $\boldsymbol{\lambda} = (-1, 1)$

2.1.3 Extended Skew-normal distribution (ESN)

Since the MSN distribution lacks the closure property under conditioning, a slight extension of this class to the so-called extended skew-normal distribution (ESN) is necessary. The ESN distribution permits the construction of multivariate skewed models that have marginal and conditional densities that are of the same form. However, the cost to be paid for gaining the latter is the loss of the χ^2 distribution of certain quadratic form (Capitanio et al., 2003). We present here the definition of the multivariate ESN distribution and from it derive the univariate equivalence. Identifiability issues of the distribution are discussed in chapter 3, and the model forms the background of what is to be used in chapter 4.

Definition 3. *A random vector* $\mathbf{Z} = (Z_1, \ldots, Z_p)'$ *has a p-dimensional ESN distri-*

bution, denoted by $\mathbf{Z} \sim ESN_p(\bar{\Omega}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda})$*, if it is continuous with PDF*

$$f(\mathbf{z}) = \frac{\phi_p(\mathbf{z}; \bar{\Omega})\Phi(\boldsymbol{\lambda}_0 + \boldsymbol{\lambda}'\mathbf{z})}{\Phi(\boldsymbol{\tau})}, \quad \mathbf{z} \in \mathbb{R}^p,$$

where $\boldsymbol{\lambda}_0 = \boldsymbol{\tau}/\sqrt{1 - \boldsymbol{\delta}'\bar{\Omega}^{-1}\boldsymbol{\delta}}$*,* $\boldsymbol{\lambda} = \bar{\Omega}^{-1}\boldsymbol{\delta}/\sqrt{1 - \boldsymbol{\delta}'\bar{\Omega}^{-1}\boldsymbol{\delta}}$*, and* $\boldsymbol{\delta} = \bar{\Omega}^{-1}\boldsymbol{\lambda}/\sqrt{1 + \boldsymbol{\lambda}'\bar{\Omega}\boldsymbol{\lambda}}$*.*

Here, $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}$ are the p-dimensional vector of shift and scale parameters respectively. For data analysis purpose, if we introduce a location-scale transformation, $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\omega}\mathbf{Z}$, where $\boldsymbol{\mu}$ and $\boldsymbol{\omega}$ are as defined in section 2.1.2, then

$$f(\mathbf{y}) = \frac{\phi_p(\mathbf{y}; \boldsymbol{\mu}, \bar{\Omega})\Phi(\boldsymbol{\lambda}_0 + \boldsymbol{\lambda}'\boldsymbol{\omega}^{-1}(\mathbf{y} - \boldsymbol{\mu}))}{\Phi(\boldsymbol{\tau})}, \quad \mathbf{y} \in \mathbb{R}^p, \tag{2.7}$$

and we write $\mathbf{Y} \sim ESN_p(\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda}_0, \boldsymbol{\lambda})$. If $p = 1$ in (2.7), we have

$$f(y; \lambda_0, \lambda_1, \mu, \sigma) = \frac{\phi\big(\frac{y-\mu}{\sigma}\big)\Phi\big(\lambda_0 + \lambda_1(\frac{y-\mu}{\sigma})\big)}{\sigma\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)}. \tag{2.8}$$

Representation (2.8) is sometimes referred to as 4-parameter skew-normal density with λ_0 & λ_1 as shift and shape parameter respectively. The moment generating function (mgf) of the above density is given by

$$M_Y(t) = \frac{\exp\Big(\mu t + \frac{\sigma^2 t^2}{2}\Big)\Phi\left(\frac{\lambda_0 + \lambda_1\sigma t}{\sqrt{1+\lambda_1^2}}\right)}{\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)}. \tag{2.9}$$

The mean and the variance of the ESN distribution is given respectively as,

$$E(Y) = \mu + \sigma\rho\Lambda(c^\star),$$

and

$$\mathrm{Var}(Y) = \sigma^2(1 - \rho^2\Lambda(c^\star)\{c^\star + \Lambda(c^\star)\}),$$

where $\Lambda = \phi/\Phi$, $\rho = \lambda_1/\sqrt{1 + \lambda_1^2}$ and $c^\star = \lambda_0/\sqrt{1 + \lambda_1^2}$. Further properties and problems of inferential procedures of this model will be discussed in chapter 3.

2.1.4 The closed skew-normal (CSN) distribution

The CSN family is constructed in the multivariate framework because it is a generalization of the multivariate skew-normal distribution such that some important properties of the normal distribution are preserved (Gonzalez-Farias et al., 2004). It is closed under marginalization, conditioning, linear transformations, sums of independent random variables from CSN family, and joint distribution of independent random variables in CSN family. We begin with a definition of the CSN distribution.

Definition 4. *Consider $p \geq 1$, $q \geq 1$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\nu} \in \mathbb{R}^q$, D an arbitrary $q \times p$ matrix, Σ and Δ positive definite matrices of dimensions $p \times p$ and $q \times q$, respectively. Then the PDF of the CSN distribution is given by:*

$$f_{p,q}(\mathbf{y}) = C\phi_p(\mathbf{y}; \boldsymbol{\mu}, \Sigma)\Phi_q(D(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \Delta), \quad \mathbf{y} \in \mathbb{R}^p, \tag{2.10}$$

with:

$$C^{-1} = \Phi_q(\mathbf{0}; \boldsymbol{\nu}, \Delta + D\Sigma D'), \tag{2.11}$$

where $\phi_p(.; \boldsymbol{\eta}, \Psi)$, $\Phi_p(.; \boldsymbol{\eta}, \Psi)$ are the PDF and CDF of a p-dimensional normal distribution with mean $\boldsymbol{\eta} \in \mathbb{R}^p$ and $p \times p$ covariance matrix Ψ. We write $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, D, \boldsymbol{\nu}, \Delta)$, if $\mathbf{y} \in \mathbb{R}^p$ is distributed as CSN distribution with parameters $q, \boldsymbol{\mu}, D, \Sigma, \boldsymbol{\nu}, \Delta$. The special case of $\boldsymbol{\nu} = 0$ in (2.10), gives,

$$f_{p,q}(\mathbf{y}) \quad = 2^q\phi_p(\mathbf{y}; \boldsymbol{\mu}, \Sigma)\Phi_q(D(\mathbf{y} - \boldsymbol{\mu}); \mathbf{0}, \Delta),$$

which is the multivariate skew-normal distribution discussed in Azzalini and Dalla Valle (1996). When $q = 1$ and $\boldsymbol{\nu} \neq 0$ in (2.10), we obtain the multivariate ESN distribution. If $p = 2$ and $q = 1$, a bivariate skew-normal distribution is derived. It is straightforward to see that the PDF in (2.10) includes the normal distribution as a special case when D and $\boldsymbol{\nu} = 0$.

The properties of CSN distributions that are required to formulate the models in chapters 4 and 5 are given below.

Properties of CSN Distribution

The CSN distribution properties of scalar multiplication, marginalization, conditioning and addition are used to construct the model described in chapter 4. The

moment generating function is used to study the extended Heckman (1979) model in chapter 5.

- The distribution function of $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, D, \boldsymbol{\nu}, \Delta)$ is given as:

$$F_{p,q}(\mathbf{y}) = C\Phi_{p+q}\left(\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}; \begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{pmatrix}, \begin{pmatrix}\Sigma & -\Sigma D'\\ -D\Sigma & \Delta + D\Sigma D'\end{pmatrix}\right), \tag{2.12}$$

  where C is as defined in (2.11).

- The distribution is closed under translation and scalar multiplications. In particular, for an arbitrary constant $\mathbf{b} \in \mathbb{R}^p$ and any real number $c \neq 0$

$$\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, D, \boldsymbol{\nu}, \Delta) \Rightarrow \mathbf{Y} + \mathbf{b} \sim CSN_{p,q}(\boldsymbol{\mu} + \mathbf{b}, \Sigma, D, \boldsymbol{\nu}, \Delta),$$

  and,

$$c\mathbf{Y} \sim CSN_{p,q}(c\boldsymbol{\mu}, \Sigma c^2, Dc^{-1}, \boldsymbol{\nu}, \Delta)$$

  In general, $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, D, \boldsymbol{\nu}, \Delta)$ if, and only if,

  $\mathbf{a}'\mathbf{Y} \sim CSN_{1,q}(\mu_{\mathbf{a}}, \Sigma_{\mathbf{a}}, D_{\mathbf{a}}, \boldsymbol{\nu}, \Delta_{\mathbf{a}})$, for every $\mathbf{a} \neq \mathbf{0}$, p-vector in $\mathbb{R}^p$, where $\mu_{\mathbf{a}} = \mathbf{a}'\boldsymbol{\mu}$, $\Sigma_{\mathbf{a}} = \mathbf{a}'\Sigma\mathbf{a}$, $D_{\mathbf{a}} = D\Sigma\mathbf{a}\Sigma_{\mathbf{a}}^{-1}$, and $\Delta_{\mathbf{a}} = \Delta + D\Sigma D' - D\Sigma\mathbf{a}\mathbf{a}'\Sigma D'\Sigma_{\mathbf{a}}^{-1}$.

- The distribution is closed under marginalization. For example, let $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, D, \boldsymbol{\nu}, \Delta)$ and partition $\mathbf{Y} = \mathbf{Y}' = (\mathbf{Y_1}', \mathbf{Y_2}')$, where $\mathbf{Y_1}$ is k dimensional, $\mathbf{Y_2}$ is $p-k$ dimensional. Then

$$\mathbf{Y_1} \sim CSN_{k,q}(\boldsymbol{\mu}_1, \Sigma_{11}, D^\star, \boldsymbol{\nu}, \Delta^\star), \tag{2.13}$$

  where $D^\star = D_1 + D_2\Sigma_{21}\Sigma_{11}^{-1}$, $\Delta^\star = \Delta + D_2\Sigma_{22.1}D_2'$, $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, and $\boldsymbol{\mu}_1$, Σ_{11}, Σ_{22}, Σ_{12}, Σ_{21} came from the corresponding partitions of $\boldsymbol{\mu}$ & Σ and D_1, D_2 from

$$D = q\begin{pmatrix} k & p-k \\ D_1 & D_2 \end{pmatrix}.$$

- The distribution is closed under the operation of conditioning.

  If $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, D, \boldsymbol{\nu}, \Delta)$, then for two subvectors $\mathbf{Y_1}$ and $\mathbf{Y_2}$, where $\mathbf{Y}' = (\mathbf{Y}_1', \mathbf{Y}_2')$, $\mathbf{Y}_1$ is k-dimensional, $1 \leq k \leq p$, and $\boldsymbol{\mu}$, Σ, D are partitioned

as above, then the conditional distribution of $\mathbf{Y}_2$ given $\mathbf{Y}_1 = \mathbf{Y}_{10}$ is

$$CSN_{p-k,q}(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{Y}_{10} - \boldsymbol{\mu}_1), \Sigma_{22.1}, D_2, \boldsymbol{\nu} - D^\star(\mathbf{Y}_{10} - \boldsymbol{\mu}_1), \Delta). \quad (2.14)$$

- The distribution is closed under sums of independent random variables. That is, if $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are independent random vectors with $\mathbf{Y}_i \sim CSN_{p,q_i}(\boldsymbol{\mu}_i, \Sigma_i, D_i, \boldsymbol{\nu}_i, \Delta_i)$, $i = 1, \ldots, n$, then

$$\sum_1^n \mathbf{Y_i} \sim CSN_{p,q^\star}(\boldsymbol{\mu}^\star, \Sigma^\star, D^\star, \boldsymbol{\nu}^\star, \Delta^\star), \quad (2.15)$$

where: $q^\star = \sum_1^n q_i$, $\boldsymbol{\mu}^\star = \sum_1^n \boldsymbol{\mu}_i$, $\Sigma^\star = \sum_1^n \Sigma_i$, $D^\star = (\Sigma_1 D_1', \ldots, \Sigma_n D_n')' \left(\sum_1^n \Sigma_i\right)^{-1}$, $\boldsymbol{\nu}^\star = (\boldsymbol{\nu}_1', \ldots, \boldsymbol{\nu}_n')'$, and:

$$\Delta^\star = \Delta^\dagger + D^\dagger \Sigma^\dagger D^{\dagger\prime} - \left[\bigoplus_1^n (D_i\Sigma_i)\right]\left(\sum_1^n \Sigma_i\right)^{-1}\left[\bigoplus_1^n (\Sigma_i D_i')\right],$$

where $\Delta^\dagger = \bigoplus_1^n \Delta_i$, $D^\dagger = \bigoplus_1^n D_i$, $\Sigma^\dagger = \bigoplus_1^n \Sigma_i$, and $\bigoplus$ is the matrix direct sum operator.

The addition of independent CSN random vectors has the dimension of p fixed but the dimension of q changes. The CSN distribution is therefore not a stable distribution.

- The moment generating function (mgf) of $\mathbf{Y}$ is given as:

$$M_{\mathbf{y}}(\mathbf{t}) = \frac{\Phi_q(D\Sigma\mathbf{t}; \boldsymbol{\nu}, \Delta + D\Sigma D')}{\Phi_q(\mathbf{0}; \boldsymbol{\nu}, \Delta + D\Sigma D')} e^{\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}, \quad \mathbf{t} \in \mathbb{R}^p. \quad (2.16)$$

The mean and the variance are respectively

$$E(\mathbf{Y}) = \frac{\partial}{\partial t} M_{\mathbf{Y}}(t)\Big|_{t=0} = \boldsymbol{\mu} + \Sigma D'\psi,$$

and

$$\begin{aligned}\mathrm{var}(Y) =& \frac{\partial^2}{\partial t \partial t'} M_{\mathbf{Y}}(t)\Big|_{t=0} - E(Y)E(Y') \\ =& \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\mu}\psi' D\Sigma + \Sigma D'\psi\boldsymbol{\mu}' + \Sigma D'\Lambda D\Sigma - E(Y)E(Y'),\end{aligned}$$

where $\psi = \frac{\Phi_q^{\star}(0;\boldsymbol{\nu},\Delta+D\Sigma D')}{\Phi_q(0;\boldsymbol{\nu},\Delta+D\Sigma D')}$ and $\Lambda = \frac{\Phi_q^{\star\star}(0;\boldsymbol{\nu},\Delta+D\Sigma D')}{\Phi_q(0;\boldsymbol{\nu},\Delta+D\Sigma D')}$ involve evaluation of first and second derivatives of multinormal integrals with respect to $\mathbf{t}$.

The CSN distribution can be represented in terms of multivariate normal and multivariate truncated normal distribution. If $\mathbf{Z} \sim N_p(\mathbf{0}, I_p)$ and $\mathbf{S} \sim N_q^{\boldsymbol{\nu}}$, that is, $\mathbf{S}$ is truncated at $\boldsymbol{\nu}$, $\mathbf{Z}$ and $\mathbf{S}$ are independent. Then the distribution of

$$\mathbf{Y} = \boldsymbol{\mu} + \Big(\Sigma^{-1} + D'\Delta^{-1}D\Big)^{-1/2}\mathbf{Z} + \Sigma D'\Big(\Delta + D\Sigma D'\Big)^{-1}\mathbf{S},$$

is $CSN_{p,q}(\boldsymbol{\mu}, \Sigma, D, \boldsymbol{\nu}, \Delta)$. Random samples can easily be simulated from the distribution using this form.

A reparametrization of the CSN distribution will result into the unified skew-normal (SUN) distribution of Arellano-Valle and Azzalini (2006). The SUN distribution unified earlier proposals extending the SN distribution, and it is a precursor to the generalization of the link between sample selection and SN distributions.

2.2 Sample selection and Skew distributions

Copas and Li's (1997) paper is probably the first instance where the link between sample selection models and skew distributions was established. Until this work, earlier appearances of the Azzalini (1985) type SN distribution, derived based on certain operations performed on the normal distribution, has been in the literature. Birnbaum (1950) in the context of educational testing showed that the SN distribution can result from linear truncation of a multivariate normal random variable. Further, Weinstein (1964) using a convolution of normal and truncated normal random variable derived a distribution similar to SN although implicitly. Roberts (1966) in the context of twin studies considered the distribution resulting from selecting the maximum/minimum value from suitably standardized measurements taken on a pair of twins. The resulting distribution is also similar to the SN distribution. In the Bayesian context, O'Hagan and Leonard (1976) suggested the use of an extended version of the SN distribution as a possible prior for a normal mean. Arnold et al. (1993) considered inference for the non-truncated marginal of a truncated bivariate normal distribution.

Other references in this category include Arnold and Beaver (2000), Arnold and Beaver (2002), Loperfido (2002), Arellano-Valle et al. (2006) and Arnold and Beaver (2007). All these revealed that simple and common nonlinear operations such as truncation, conditioning and censoring carried out on normal random variables lead invariably to versions of skew-normal random variables.

Arellano-Valle et al. (2002) and Arellano-Valle and Azzalini (2006) put forward a formula for the derivation of Azzalini (1985) type SN distribution using a conditioning approach. This was extended in Arellano-Valle et al. (2006) to establish a link between sample selection and SN distributions. The model, which is simply a conditional distribution, is defined as follows.

Definition 5. *Let* $\mathbf{S}^\star \in \mathbb{R}^q$ *and* $\mathbf{Y}^\star \in \mathbb{R}^p$ *be two random vectors, and denote by* $\mathbf{C}$ *a measurable subset of* $\mathbb{R}^q$*. A selection distribution is defined as the conditional distribution of* $\mathbf{Y}^\star$ *given* $\mathbf{S}^\star \in \mathbf{C}$ *(i.e.* $\mathbf{Y}^\star|\mathbf{S}^\star \in \mathbf{C}$*). A random vector* $\mathbf{Y} \in \mathbb{R}^p$ *is said to have a selection distribution if* $\mathbf{Y} \stackrel{d}{=} (\mathbf{Y}^\star|\mathbf{S}^\star \in \mathbf{C})$*.*

If $\mathbf{C} = \mathbb{R}^q$, then there is no selection. The model can be viewed as a truncated distribution when $\mathbf{Y}^\star = \mathbf{S}^\star$. In particular, if $\mathbf{Y}^\star$ in definition 5 has PDF $f_{\mathbf{Y}^\star}$ say, then $\mathbf{Y}$ has a PDF f_Y given by

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}^\star}(\mathbf{y}) \frac{P(\mathbf{S}^\star \in \mathbf{C} \,|\, \mathbf{Y}^\star = y)}{P(\mathbf{S}^\star \in \mathbf{C})}.$$

Selection distributions depend on the subset $\mathbf{C}$ of $\mathbb{R}^q$. The usual selection subset is defined by

$$\mathbf{C}(\beta) = \{\mathbf{s} \in \mathbb{R}^q \,|\, \mathbf{s} > \beta\},$$

where β is a vector of truncation levels. A hidden truncation equivalence of selection distributions consist of upper and lower truncation subset defined by

$$\mathbf{C}(\alpha, \beta) = \{\mathbf{s} \in \mathbb{R}^q \,|\, \alpha > \mathbf{s} > \beta\}.$$

A special case of this subset with $p = q = 1$ is considered in Arnold et al. (1993). For this thesis, we will focus on the subset $\mathbf{C}(\mathbf{0})$ which leads to simple selection distribution. Note that the only difference between using $\mathbf{C}(\beta)$ and $\mathbf{C}(\mathbf{0})$ is essentially a location change, since no symmetry around $\mathbf{0}$ is assumed. In this case, the distribution $\mathbf{X} = (\mathbf{Y}^\star|\mathbf{S}^\star > \mathbf{0})$ can be written as

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}^\star}(\mathbf{y}) \frac{P(\mathbf{S}^\star > \mathbf{0} \,|\, \mathbf{Y}^\star = y)}{P(\mathbf{S}^\star > \mathbf{0})}. \tag{2.17}$$

To illustrate how (2.17) is linked with skew-distributions, consider a multivariate extension of Copas and Li (1997) model.

$$\begin{aligned} \mathbf{Y}^\star &= \boldsymbol{\mu} + \boldsymbol{\varepsilon}_1, \quad \boldsymbol{\varepsilon}_1 \sim N_p(\mathbf{0}, \Sigma) \\ \mathbf{S}^\star &= -\boldsymbol{\nu} + D\boldsymbol{\mu} + \boldsymbol{\varepsilon}_2, , \quad \boldsymbol{\varepsilon}_2 \sim N_q(\mathbf{0}, \Delta), \end{aligned} \tag{2.18}$$

where $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are independent random vectors, and $D(q \times p)$ is an arbitrary matrix, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\nu} \in \mathbb{R}^q$, and $\Delta(q \times q) > 0$. The joint distribution of $\mathbf{Y}^\star$ and $\mathbf{S}^\star$ is:

$$\begin{pmatrix} \mathbf{Y}^\star \\ \mathbf{S}^\star \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} \boldsymbol{\mu} \\ -\boldsymbol{\nu} \end{pmatrix}, \begin{pmatrix} \Sigma & \Sigma D' \\ D\Sigma & \Delta + D\Sigma D' \end{pmatrix} \right).$$

But the conditional density $(\mathbf{y}^\star | \mathbf{s}^\star > \mathbf{0})$ can easily be written as in equation (2.17), which simplifies to,

$$f_{\mathbf{Y}^\star|\mathbf{S}^\star>\mathbf{0}}(\mathbf{y}^\star|\mathbf{s}^\star > \mathbf{0}) = C\phi_p(\mathbf{y}; \boldsymbol{\mu}, \Sigma)\Phi_q(D(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \Delta), \tag{2.19}$$

where C is as defined in (2.11). This is a CSN distribution. A similar argument can be used to show that the univariate Copas and Li (1997) model is essentially the extended skew-normal distribution given in (2.8).

2.3 Other families of Skew distributions

Apart from the Azzalini (1985) type skew-symmetric distributions, which are constructed by perturbation of symmetric PDFs, other methods for the construction of skew distributions have been studied. An example of skew distribution constructed with different scale factors is studied in Fernandez and Steel (1998) and Ferreira and Steel (2007). Other methods include derivation of skew distributions from distributions of order statistics (e.g. Jones (2004)), and skew distributions obtained via the transformation approach (e.g. Jones and Pewsey (2009)). We will use the skew distribution based on the latter in a copula based sample selection model in chapter 6.

2.4 Motivating Example-The MINT Trial

The data set used to illustrate the methods proposed in the first part of this thesis is presented in this section. The data set is obtained from a two-arm clinical trial in patients suffering from neck disability called MINT study. This data is used to illustrate the proposed methods in chapters 3-6 of this thesis.

MINT is a multi-center randomized controlled trial to estimate the clinical effectiveness of a stepped care approach to whiplash injuries on clinical outcomes over 12 months, the effectiveness in pre-specified sub-groups of patients (those with severe physical symptoms, prior neck problems, psychological or physical risk factors for poor outcome, and those seeking compensation), and the costs and cost-

effectiveness of each strategy (Lamb et al. (2007)). Treating patients at the lowest appropriate treatment tiers, and only stepping up to more intensive treatment as clinically required for a neck injury caused by a sudden forward movement of the upper body is called a stepped-care approach to whiplash injury. The trial is a two-stage randomized controlled trial to evaluate two stepped care evaluation methods. These are:

- The Whiplash book
- Physiotherapy.

Consequently, the first stage randomization (with about four thousand participants), which was at a cluster level, was done when the patients first attended the emergency departments of the hospitals used in the study. Thus, a comparison between the use of 'The Whiplash Book' (Burton et al., 2001) *versus* 'Usual Advice' was done at this level. The second stage randomization is an individually randomized trial of physiotherapy *versus* reinforcement of advice given in Emergency Department. The main eligibility criteria for entry to Stage 2 was that the patients have no contra-indications to physiotherapy treatment and report symptoms in the 24 hours before attendance at the physiotherapy research clinic approximately three weeks after attendance at ED. Details of randomization and data collection methods for Stage 2 MINT trial are given in Lamb et al. (2007). We present some attributes of the data set in Stage 2 of MINT trial.

Stage 2 Physiotherapy *versus* Reinforcement of Advice

Six hundred patients were randomized into either physiotherapy or reinforcement of advice. It was expected that all treatments would be completed within four months of the patient's first attendance at emergency department. The following treatments are included in the physiotherapy package (Lamb et al., 2007):

1. Mobilization (gentle manipulation) of the cervical and upper thoracic spine.
2. Exercises for the cervical spine, thoracic spine and shoulder to improve range of movement and muscle control.
3. A cognitive behavioral approach to treatment delivery, which has been effective in physiotherapy for other painful conditions.

For advice reinforcement, patients receive a single 40-minute session of advice from a physiotherapist. Details of the four outcome measures are given in Lamb et al.

(2007). Our main focus will be on the primary outcome of interest which is return to normal function after the whiplash injury, measured using the NDI scores. The NDI is a self-completed questionnaire which assess pain-related activity restrictions in 10 areas including personal care, lifting, sleeping, driving, concentration, reading and work. It was developed in 1989 by Howard Vernon as a modification of the Oswestry Low Back Pain Disability Index. The NDI has been shown to be reliable and valid (Vernon and Mior, 1991), hence its use as a standard instrument for measuring self-rated disability due to neck pain by clinicians and researchers.

Each of the 10 items on the questionnaire is scored from 0-5. In effect, the maximum obtainable score is 50. Some respondents will not complete all the questions (called item non-response in surveys). The average of all other items is scaled to give an imputed score if one or two items are missing. The scoring intervals are interpreted as follows:

- $0 - 4$ = No Disability
- $5 - 14$ = Mild Disability
- $15 - 24$ = Moderate Disability
- $25 - 34$ = Severe Disability
- $35 - 50$ = Complete Disability.

Measurements were taken at baseline, and at four months interval for a complete calendar year (0, 4, 8 and 12 months). Exploration of salient variables and other interesting features of the data are examined and are presented in next section.

Numerical Exploration of MINT's Data

There are 599 patients with a total of 1934 measurements and 342 patients have complete observations (i.e. scores at all measurements occasion). Further, approximately 50% of the patients are in the two treatment groups resulting in balanced randomization in terms of patients number.

Table 2.1 shows the number of questions missing at various time points. It is observed that question 8 (question related to driving) recorded the highest number of missing observations while question 4 (question on reading) was answered by most patients. The driving question consistently recorded high missing value across all the four measurement occasions. Analogous to most longitudinal studies, the number of missing scores (Table 2.2) at the last measurement occasion, month 12,

Table 2.1: Missingness per question during the trial; 599 patients.

Time	$q1$	$q2$	$q3$	$q4$	$q5$	$q6$	$q7$	$q8$	$q9$	$q10$
Baseline	2	3	2	0	1	0	4	43	0	1
Month 4	98	97	100	96	97	96	98	120	96	96
Month 8	106	104	108	103	106	104	108	131	104	104
Month 12	123	122	125	120	122	120	123	143	121	121
Total	329	326	335	319	326	320	333	**437**[a]	321	322

[a]Driving question with highest number of non-response.

Table 2.2: Scoring Interval and Overall missingness with Measurement time.

	Scoring Interval					missingness	
Time	0-4	5-14	15-24	25-34	35-50	num	%
	Phy. Adv.	Phy. Adv.	Phy. Adv.	Phy. Adv.	Phy. Adv.	Phy. Adv.	Phy. Adv.
Base	1 2	55 72	121 131	76 56	22 12	25 26	4.2 4.3
M4	33 32	104 104	62 69	25 25	8 7	68 62	11.4 10.4
M8	56 26	96 92	56 51	17 15	5 3	70 76	11.7 12.7
M12	70 80	84 87	51 45	12 12	5 1	78 74	13.0 12.4

was highest. Only six patients reported complete disability (35-50 scores on the NDI scale) at the last measurement occasion with 45.9% reported to have no disability (see Table 2.2, and scoring interval, Page 21). This result is obviously as expected when subjective endpoints are accessed in clinical trials. In addition, there is wide variability in patients' age distribution. The mean age is approximately 41 years with range 18 to 78 years respectively. The mean age of patients in 'Usual Advice' and 'Physio' treatment is 40.8 and 41.2 respectively.

Assessing Normality of the Observed scores

Since scores are formed by adding up items on a scale, the observed NDI scores are inevitably skewed (see Figure 2.3). A chi-square (also known as gamma plot, see Johnson and Wichern (2007)) is used to assess item normality of the NDI scale. Figure 2.4 shows the chi-square plot for measurements at baseline and the three follow-up. There is obvious departure from straight line through the origin. The departure became more pronounced as follow-up increases with measurements at month 12 having the greatest departure. This could be due to the fact that more patients drop out at month 12 than any other follow-up period. Thus, the observed scores represent a selected population hence skewed. We further corroborate this conclusion by the use of the multivariate extension of Shapiro-Wilk test (mvnormtest

in R) with all measurements occasion reporting a significant p-value thereby rejecting the Null Hypothesis of multivariate item normality for the NDI scale.

Assessing Normality of the Residuals

In longitudinal studies, the usual assumption for modeling observed responses at the measurement occasions is that the residuals follow joint multivariate normality. Often, this assumption is not realistic. Figure 2.5 shows the q-q plots of the residuals obtained after fitting univariate normal error regression models to the observed scores at baseline, 4, 8 and 12 months follow-up. The plots deviate from the 'straightness' that is required to confirm normality with the heaviest deviation at month 8. A formal test using the correlation coefficients 0.993, 0.980, 0.972 and 0.973 for baseline, month 4, month 8 and month 12 respectively showed that the normality assumption is rejected when compared with the critical value (0.9953) corresponding to the data at hand at 5% level of significance. Indeed, the normality assumption is rejected marginally for the four measurements occasions. This, in principle, implies that conditional normality is not tenable for any of the measurement occasion in this data set and models to be used must accommodate skewness to avoid wrong inferences.

2.5 Concepts of Missing Data

As shown above, the NDI scores is incomplete both at the unit and item levels. Similarly, the bladder cancer data that will be used in part II of this thesis suffers from some form of missing data problem. The incompleteness of the data sets may lead to results that are different from those that would have been obtained had the data sets been completely observed. Hence, it is important to handle missingness carefully. In this section, we introduce notation and fundamental concepts that are used in the area of incomplete data.

Notation for Missing Data

We follow the standard notion for missing data due to Rubin (1976) and used by Verbeke and Molenberghs (2000). Suppose that for subject i, $i = 1, 2, ...N$, a sequence of measurements Y_{ij} is designed to be measured at time points t_{ij}, $j = 1, 2, ...n_i$. The outcome vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$ that would have been recorded if there had been no missing data is referred to as the *complete* data. Suppose further that, for each measurement in the series, a corresponding missingness indicator R_{ij} is defined as:

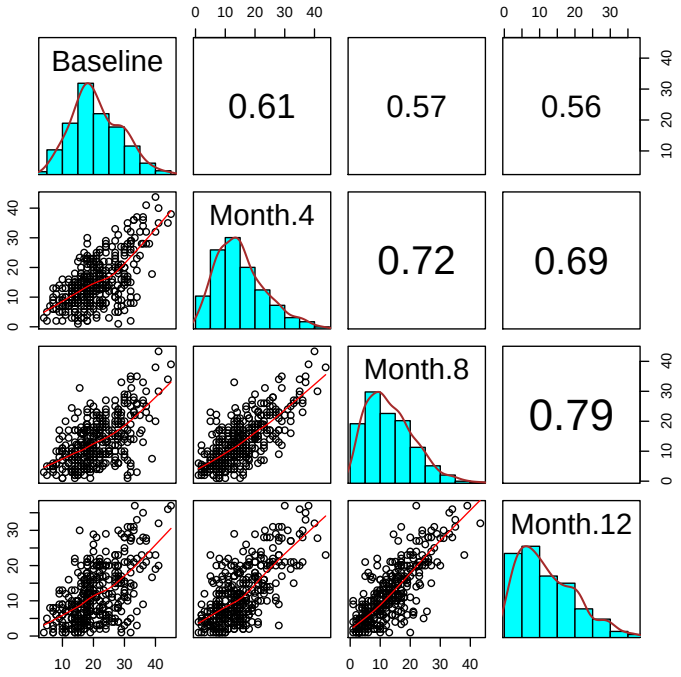

Figure 2.3: Marginal distributions and Correlations at Baseline, Month 4, 8 and 12 for the NDI scores

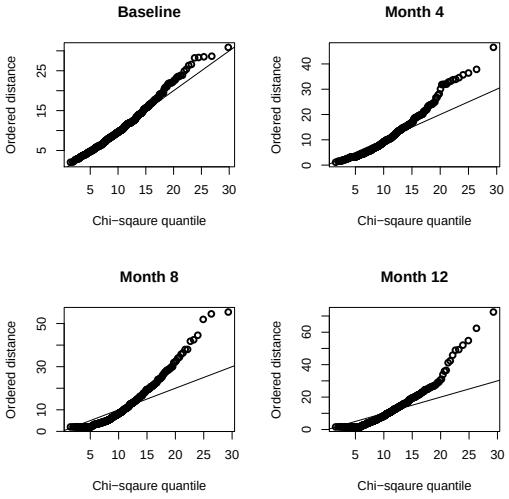

Figure 2.4: Chi-square plots for items at baseline, month 4, month 8 and month 12

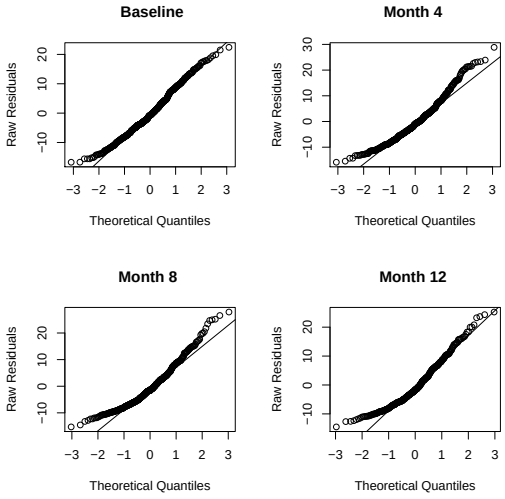

Figure 2.5: Q-Q plots for residuals of scores at baseline, month 4, month 8 and month 12

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

which are grouped into a vector $\mathbf{R}_i$ of the same length as $\mathbf{Y}_i$. The set of measurements, along with the missingness indicators, $(\mathbf{Y}_i, \mathbf{R}_i)$, is referred to as the *full* data. Typically, $\mathbf{Y}_i$ can be partitioned into two sub-vectors: $\mathbf{Y}_i^{obs}$ consisting of those Y_{ij} for which R_{ij}=1, and $\mathbf{Y}_i^{miss}$ consisting of the remaining components, which are referred to as observed and missing components respectively.

Modeling Framework of Missing Data

Models for incomplete cross sectional or longitudinal data involves working with the joint density $f(y_i, r_i|X_i, Z_i, \theta, \psi)$ where X_i and Z_i are design matrices for fixed and random effects, respectively, and θ and ψ are respective parameter vectors describing the response and missingness process. The use of joint density is prompted by the presence of the two stochastic components $\mathbf{Y}_i$ and $\mathbf{R}_i$. The modeling frameworks have been elucidated in statistical literatures and it is based on the choice of the factorization of the joint density above.

Since the patients are considered independent, the joint density (after suppressing dependence on X_i and Z_i) can be factored as either

$$f(y_i, r_i|X_i, Z_i, \theta, \psi) = f(y_i|\theta) f(r_i|y_i, \psi), \tag{2.20}$$

$$f(y_i, r_i|X_i, Z_i, \theta, \psi) = f(y_i|r_i, \theta) f(r_i|\psi) \tag{2.21}$$

or as

$$f(y_i, r_i|X_i, Z_i, \theta, \psi) = f(y_i|b_i, \theta) f(r_i|b_i, \psi), \tag{2.22}$$

where in (2.22) the response and missingness processes are independent conditional on a common set b_i of latent variables or random effects.

The factorization in (2.20) is termed a selection model (Rubin, 1976). This model is often an obvious choice in clinical trials. In trials context, incomplete data is often dependent on treatment response. This implies that patients are selected for missingness by their response. The factorization in (2.21) is termed pattern-mixture models (Little, 1993). In this case, different patterns of response can be proposed for patients who have or do not have missing values. The third factorization (2.22) is termed shared-parameter models. Details of this modeling framework can be found

in Wu and Carroll (1988) and Wu and Bailey (1989).

The comparison of parameter of interest (θ) is not immediately possible in the three modeling frameworks. Clearly, θ in (2.20) represents marginal effects whereas θ in (2.21) and (2.22) describe conditional effects. Any attempt to obtain marginal effects will require marginalization over the missingness pattern for the former or over the random effects for the latter.

Missing Data Mechanisms

The basic taxonomy for classifying missingness process was developed in the selection model framework (Rubin, 1976). The precise form of the second term in the right hand side of (2.20) which can be expressed as $f(r_i|y_i, \psi) = f(r_i|y_i^{obs}, y_i^{miss}, \psi)$ defines the missingness mechanism. In line with Diggle and Kenward (1994) and Little and Rubin (2002), the missing data mechanisms are described below.

The data is said to be missing completely at random (MCAR) if missingness does not depend on either the observed or the unobserved responses. Mathematically,

$$f(r_i|y_i^{obs}, y_i^{miss}, \psi) = f(r_i|\psi). \tag{2.23}$$

In Little (1995), covariate dependent missingness are classified as MCAR missingness. This was further stressed in Carpenter et al. (2002). If missingness depends on those values of y_i that are observed and not on the unobserved components, the data are said to be missing at random (MAR). Mathematically,

$$f(r_i|y_i^{obs}, y_i^{miss}, \psi) = f(r_i|y_i^{obs}, \psi). \tag{2.24}$$

This missingness assumption is less restrictive than MCAR.

Finally, if the missingness depends on unobserved components of y_i i.e y_i^{miss} then the data is missing not at random (MNAR). In this case, we cannot simplify $f(r_i|y_i^{obs}, y_i^{miss}, \psi)$.

Importantly, it should be noted that MCAR, MAR and MNAR are assumptions made regarding the underlying missingness process, therefore absolute certainty about them cannot be guaranteed. Indeed, the validity of inferences made under different statistical methods depends on the assumption made about the missingness process. Since MNAR missingness cannot be ruled out in practice, the principal focus of this thesis is to develop models in a sample selection framework (which is a form of MNAR missingness), but with more flexible underlying distributional assumption.

According to Carpenter et al. (2002), four different approach to the analysis of missing data can be distinguished:

- Perform the analysis only on those subjects who complete the trial;
- Analyse only the available data;
- Use a single or multiple imputation technique to replace the missing observations with plausible values, then analyse the complete data set(s); and
- Model observed data and the missingness process jointly.

The first option yields a complete case analysis and form the basis of the discussion in chapter 3. The second option is the likelihood-based approach of using available information only. Single and multiple imputation techniques has been well established in the literature (Rubin (1987), Rubin (1996), Schafer (1997), Schafer (1999), Little and Rubin (2002)). Chapter 7 of this thesis is devoted to the use of multiple imputation in recurrent event data with dropouts. The fourth option is usually the most complex, and also the most useful as it gives room to easily assess subtle assumptions behind other methods in a sensitivity analysis framework. Chapters 4-6 of this thesis is devoted to this method.

Chapter 3

Ignorable Missing Data Methods and Sample Selection

We noted in chapter 2 the link between sample selection and skew distributions, and that the hidden truncation models can be considered a special case of sample selection models. Due to this link, it would be logical to use skew distributions to model data arising from hidden truncation or sample selection. In this chapter, we consider complete case analysis for data arising from sample selection with underlying normal and skew-normal distributions. The performances of the Azzalini skew-normal distribution, the extended skew-normal distribution, and a new class of model, which we refer to as an extended two-parameter generalized skew-normal (EGSN) distribution are evaluated in a simulation study. Since the scores are bounded, we also consider modeling the outcome using doubly truncated skew-normal distribution.

3.1 Copas and Li (1997) Sample selection model

Consider a univariate case of the model given in equation (2.18), but with error distributions unspecified for the moment. That is, let $Y_i^\star$ be the outcome variable of interest, assumed linearly related to covariates x_i through the standard multiple regression

$$Y_i^\star = \beta' x_i + \sigma\varepsilon_{1i}, \quad i = 1, \ldots, N.$$

Suppose the main model is supplemented by a selection (missingness) equation

$$S_i^\star = \gamma' x_i + \varepsilon_{2i}, \quad i = 1, \ldots, N$$

where β and γ are unknown parameters and x_i are fixed observed charac-

teristics not subject to missingness, the variance of $S_i^\star$ is fixed as 1 because the variance is not identifiable from sign alone. Selection is modeled by observing $Y_i^\star$ only when $S_i^\star > 0$ (the 0 threshold is arbitrary since no symmetry is assumed), i.e. we observe $S_i = I(S_i^\star > 0)$ and $Y_i = Y_i^\star S_i$ for $n = \sum_{i=1}^{N} S_i$ of N individuals. Thus an observation has the conditional density

$$f(y|x, S^\star > 0) = \frac{f(y, S^\star > 0|x)}{P(S^\star > 0|x)} = \frac{f(y|x)P(S^\star > 0|y, x)}{P(S^\star > 0|x)}. \tag{3.1}$$

Equation (3.1) is the univariate case of (2.17). The quantity $f(y|x)$ is a proper PDF, with a skewing function $P(S^\star > 0|y, x)$, and a normalizing function $P(S^\star > 0|x)$. It is straightforward to show that under the additional assumption

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\};$$

$$f(y|x, S = 1; \Theta) = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\beta' x}{\sigma}\right)\Phi\left(\frac{\gamma' x + \rho\left(\frac{y-\beta' x}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)}{\Phi(\gamma' x)}, \tag{3.2}$$

(see Copas and Li (1997)), where $\Theta = (\beta, \sigma, \gamma, \rho)$. The parameter $\rho \in$ [-1,1] determines the correlation of $Y_i^\star$ and $S_i^\star$, and hence the severity of the selection process.

Model (3.2) includes the three missing data mechanisms discussed in section 2.5. If the non-intercept terms in γ, as well as ρ are 0 in (3.2), the data are MCAR. If $\rho = 0$ in (3.2) the data are MAR, and valid inference about the conditional distribution of Y given x can be made when adjustment for missing data are done using covariates on complete cases. If $\rho \neq 0$ in (3.2), then the missing data are MNAR. In this case, the missing data process is said to be informative or non-ignorable, as valid inference depends on adequate adjustment for the selection process.

As expected, from Arellano-Valle et al. (2006), equation 3.2 belongs to the extended skew-normal distribution family. To see this, we let $\mu = \beta' x$, $\lambda_0 = \gamma' x/\sqrt{1-\rho^2} \in \mathbb{R}$ and $\lambda_1 = \rho/\sqrt{1-\rho^2} \in \mathbb{R}$ in (3.2); we then have the PDF written in the four-parameter ESN form given in equation (2.8).

In principle, (3.2) can also be derived using hidden truncation methods. In line with Arnold et al. (1993), the non-truncated marginal of a truncated bivariate normal distribution is essentially an ESN distribution. In particular, suppose Z and S are two independent random variables, with arbitrary and possibly different distributions, and the outcome Z is observed only if S satisfies the constraints $\lambda_0 + \lambda_1 Z > S$. If we further assume that Z has density function ψ_1 with associated

distribution function Ψ_1 and S has density function ψ_2 with distribution function Ψ_2, then the conditional density of $Z|(\lambda_0 + \lambda_1 Z > S)$, according to Arnold and Beaver (2002), is

$$f(z|(\lambda_0 + \lambda_1 Z > S) = \frac{\psi_1(z)\Psi_2(\lambda_0 + \lambda_1 z)}{P(\lambda_0 + \lambda_1 Z > S)}. \tag{3.3}$$

In particular, if Z and S are normally distributed in (3.3), the resulting distribution is a two-parameter ESN distribution, with density given by

$$f(z|(\lambda_0 + \lambda_1 Z > S) = \frac{\phi(z)\Phi(\lambda_0 + \lambda_1 z)}{\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)},$$

which becomes (2.8) after a location-scale transformation $Y = \mu + \sigma Z$.

In general, as equation (2.17) shows, it is straightforward to establish a link between sample selection and families of extended skew-elliptical distributions. The Copas and Li (1997) model used underlying bivariate normal distribution which results in the extended skew-normal distribution as we have shown here. Marchenko and Genton (2012) used underlying student's-t distribution, and established a link with the extended skew-t distribution. The use of skew-elliptical distributions to model complete cases may therefore appear to be a good practice in the sample selection framework. We examine the pros and cons of regression models using ESN distribution next.

3.2 Regression models with ESN error distribution

Suppose $Y_1, \ldots, Y_n$ are independent realization from Y with covariates $x_1, \ldots, x_n$, the model can be written as

$$Y_i = \beta' x_i + \sigma\varepsilon_i, \quad \varepsilon_i \sim \text{ESN}(\mu^\star, \sigma^{2\star}, \lambda_0, \lambda_1),$$

where $\mu^\star = \sigma\rho\Lambda(c^\star)$ and $\sigma^{2\star} = \sigma^2(1 - \rho^2\Lambda(c^\star)\{c^\star + \Lambda(c^\star)\})$ and Λ, ρ and $c^\star$ are as defined in section 2.1.3. Unlike the normal errors, these errors have non-zero conditional mean. The amount of the bias is given by

$$E(Y_i - \beta' x_i) = \sigma\rho\Lambda(c^\star), \tag{3.4}$$

where $\Lambda(.)$ is the inverse Mills ratio. The MLEs of the 4-parameter ESN model are obtained by simultaneously maximizing the log-likelihood function given below.

$$l(\Theta) = \frac{-n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\beta x_i)^2}{\sigma^2} + \sum_{i=1}^{n}\ln\Big[\Phi\Big(\lambda_0+\lambda_1\Big(\frac{y_i-\beta x_i}{\sigma}\Big)\Big)\Big] \\ -n\ln\Big[\Phi\Big(\lambda_0(1+\lambda_1)^{-1/2}\Big)\Big],$$

where $\Theta = (\mu, \sigma, \lambda_0, \lambda_1)$.

The ESN model suffers from severe identifiability problems and, as such, Arnold et al. (1993) suggested the use of profile likelihood to help study the uncertainty in the MLEs. The reason given for this is that the distribution can be uninformative about all of the population parameters, even with large sample size. In particular, the model may be unidentifiable in the sense that for some $(\lambda_0, \lambda_1) \neq (\lambda_0^*, \lambda_1^*)$, $f(y; \lambda_0, \lambda_1) = f(y; \lambda_0^*, \lambda_1^*)$. For example, regardless of the value of λ_0, the ESN distribution reduces to the normal distribution when $\lambda_1 = 0$. In addition, like in Azzalini SN distribution where say, $SN(9)$ and $SN(10)$ are indistinguishable, (λ_0, λ_1) & $(\lambda_0^*, \lambda_1^*)$ may also be indistinguishable. To see this, suppose for a given density with parameters $(\lambda_0, \lambda_1) = \theta_1$ and for a given $\epsilon > 0$ there is another pair of parameter $(\lambda_0^*, \lambda_1^*) = \theta_2$ such that

$$\Delta(\theta_1, \theta_2) = \max|f(z; \lambda_0, \lambda_1) - f(z; \lambda_0^*, \lambda_1^*)| < \epsilon,$$

then θ_1 & θ_2 are indistinguishable. Examples of such 'equal' models include, $\Delta((3,3),(2,3)) = 0.02$ and $\Delta((3,2),(2,1.3)) = 0.01$. The smaller the value of ϵ, the less the two models are distinguishable (see Figure 3.1 for the plot of the latter parameter combination).

3.3 Generalized Skew-normal distribution

One of the generalization of the Azzalini (1985) SN distribution is the two-parameter generalized skew-normal (GSN) distribution introduced by Jamalizadeh et al. (2008). Its PDF was given as

$$f(z; \lambda_1, \lambda_2) = \frac{2\pi}{\cos^{-1}\Big(\frac{-\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda_2^2}}\Big)}\phi(z)\Phi(\lambda_1 z)\Phi(\lambda_2 z), \quad z \in \mathbb{R}. \tag{3.5}$$

The author realized in their follow-up papers (Jamalizadeh and Balakrishnan, 2009, 2010) that the distribution is in fact special cases of the multivariate unified skew-normal (SUN) presented by Arellano-Valle et al. (2006), which in itself is a reparametriza-

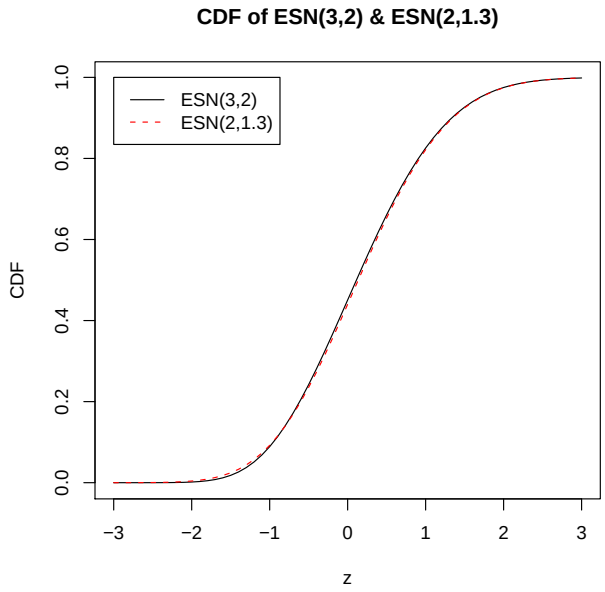

Figure 3.1: Two indistinguishable parameter combination for two-parameter ESN, $\Delta((3,2),(2,1.3)) = 0.01$

tion of the CSN distribution of Gonzalez-Farias et al. (2004). They gave basic properties of the distribution, but technical properties can easily be derived if one takes advantage of the CSN distribution reparametrization. For example, the addition of an independent random variable from $\text{GSN}(\lambda_1, \lambda_2)$ and normal $N(0,1)$ random variable is still in the two-parameter generalized skew-normal distribution as the following theorem shows.

Theorem 2. *Let* $\mathbf{Y} \sim CSN_{p,q}(\boldsymbol{\mu}, \Sigma, D, \boldsymbol{\nu}, \Delta)$*, with parameters as defined in* (2.10)*. Let also* $\mathbf{X} \sim N_p(\boldsymbol{\mu}_x, \Sigma_x)$*,* $\Sigma_x > 0$ *be independent of* $\mathbf{Y}$*, then*

$$\mathbf{Y} + \mathbf{X} \sim CSN_{p,q}\Big(\boldsymbol{\mu} + \boldsymbol{\mu}_x, \Sigma + \Sigma_x, D\Sigma(\Sigma + \Sigma_x)^{-1}, \boldsymbol{\nu}, \Delta + (D(I - \Sigma(\Sigma + \Sigma_x)^{-1}))\Sigma D'\Big).$$

If we apply the theorem to $Z_{\lambda_1,\lambda_2} \sim \text{GSN}(\lambda_1, \lambda_2)$ and $X \sim N(0,1)$, that is,

$$Z_{\lambda_1,\lambda_2} \sim CSN_{1,2}\Big(0, 1, (\lambda_1, \lambda_2)', (0,0)', \Delta = I_2\Big),$$

where I_2 is a 2×2 identity matrix. Then,

$$\Big(Z_{\lambda_1,\lambda_2} + X\Big) \sim CSN_{1,2}\left[0, 2, (\lambda_1/2, \lambda_2/2)', (0,0)', \begin{pmatrix} 1 + \lambda_1^2/2 & \lambda_1\lambda_2/2 \\ \lambda_1\lambda_2/2 & 1 + \lambda_2^2/2 \end{pmatrix}\right].$$

By using scalar multiplication properties of the CSN distribution, we have

$$\frac{1}{\sqrt{2}}\Big(Z_{\lambda_1,\lambda_2}+X\Big)\sim CSN_{1,2}\Bigg[0,1,(\lambda_1/\sqrt{2},\lambda_2/\sqrt{2})',(0,0)',\begin{pmatrix}1+\lambda_1^2/2 & \lambda_1\lambda_2/2\\ \lambda_1\lambda_2/2 & 1+\lambda_2^2/2\end{pmatrix}\Bigg],$$

which is a two-parameter generalized skew-normal distribution with parameters $(\lambda_1/\sqrt{2},\lambda_2/\sqrt{2})$. This theorem is a special case of the general method of adding independent random variables from the CSN distribution given in (2.15).

We now construct two classes of three-parameter extensions of the Jamalizadeh et al. (2008) model. The first extension is written as a special case of the CSN distribution and the second extension adds a shift parameter to the Jamalizadeh et al. (2008) model.

3.3.1 A three-parameter generalized skew-normal distribution

We use the CSN distribution given in chapter 2 to define a three-parameter generalized skew-normal distribution, GSN$(\lambda_1,\lambda_2,\lambda_3)$.

Definition 6. *A random variable $Z_{\lambda_1,\lambda_2,\lambda_3}$ is said to have a three-parameter generalized skew-normal distribution if its PDF can be written as*

$$f(z;\lambda_1,\lambda_2,\lambda_3)=\frac{1}{\Phi_3(\mathbf{0};\rho_{12},\rho_{13},\rho_{23})}\phi(z)\Phi_3((\lambda_1,\lambda_2,\lambda_3)'z;\mathbf{0},I_3),\qquad(3.6)$$

where $\rho_{12}=\lambda_1\lambda_2/\sqrt{1+\lambda_1^2}\sqrt{1+\lambda_2^2}$, $\rho_{13}=\lambda_1\lambda_3/\sqrt{1+\lambda_1^2}\sqrt{1+\lambda_3^2}$, ρ_{23}=$\lambda_2\lambda_3/\sqrt{1+\lambda_2^2}\sqrt{1+\lambda_3^2}$, *and I_3 is a 3×3 identity matrix.*

We write $Z\sim CSN_{1,3}(0,1,D=(\lambda_1,\lambda_2,\lambda_3)',\nu=(0,0,0)',I_3)$. Since the PDF given by (3.6) is in a CSN form, it is trivial to show that it is a proper PDF.

In order to avoid the evaluation of three dimensional integral present in the CSN representation (3.6), one can re-write the expression in the form given by Jamalizadeh et al. (2008). To do this, we consider the following lemma.

Lemma 1. *If* $R=\begin{pmatrix}1 & \rho_{12} & \rho_{13}\\ \rho_{12} & 1 & \rho_{23}\\ \rho_{13} & \rho_{23} & 1\end{pmatrix}$, *then*

$$\Phi_3(\mathbf{0};R)=\frac{2\pi-\cos^{-1}(\rho_{12})-\cos^{-1}(\rho_{13})-\cos^{-1}(\rho_{23})}{4\pi}.$$

Thus equation (3.6) can be written as

$$\frac{4\pi}{2\pi-\cos^{-1}(\rho_{12})-\cos^{-1}(\rho_{13})-\cos^{-1}(\rho_{23})}\phi(z)\Phi(\lambda_1 z)\Phi(\lambda_2 z)\Phi(\lambda_3 z),$$

and we write $Z_{\lambda_1,\lambda_2,\lambda_3} \sim \text{GSN}(\lambda_1, \lambda_2, \lambda_3)$. We denote

$$\frac{4\pi}{2\pi - \cos^{-1}(\rho_{12}) - \cos^{-1}(\rho_{13}) - \cos^{-1}(\rho_{23})} = \frac{1}{\Phi_3(\mathbf{0}; \rho_{12}, \rho_{13}, \rho_{23})} = K(\lambda_1, \lambda_2, \lambda_3).$$

Basic properties of GSN$(\lambda_1, \lambda_2, \lambda_3)$

Using the properties of CSN distribution, simple properties of the $\text{GSN}(\lambda_1, \lambda_2, \lambda_3)$ distribution can be obtained.

1. $\text{GSN}(\lambda_1, \lambda_2, 0) = \text{GSN}(\lambda_1, 0, \lambda_2) = \text{GSN}(0, \lambda_1, \lambda_2) = \text{GSN}(\lambda_1, \lambda_2)$
2. $\text{GSN}(\lambda, 0, 0) = \text{GSN}(0, \lambda, 0) = \text{GSN}(0, 0, \lambda) = \text{SN}(\lambda)$
3. GSN(0,0,0) = N(0,1)
4. $Z \sim \text{GSN}(\lambda_1, \lambda_2, \lambda_3)$, then $-Z \sim \text{GSN}(-\lambda_1, -\lambda_2, -\lambda_3)$
5. The distribution function of $Z \sim \text{GSN}(\lambda_1, \lambda_2, \lambda_3)$ is

$$K(\lambda_1, \lambda_2, \lambda_3)\Phi_4\left(\begin{pmatrix} Z \\ 0 \\ 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -\lambda_1 & -\lambda_2 & -\lambda_3 \\ -\lambda_1 & 1+\lambda_1^2 & \lambda_1\lambda_2 & \lambda_1\lambda_3 \\ -\lambda_2 & \lambda_1\lambda_2 & 1+\lambda_2^2 & \lambda_2\lambda_3 \\ -\lambda_3 & \lambda_1\lambda_3 & \lambda_2\lambda_3 & 1+\lambda_3^2 \end{pmatrix}\right).$$

Figure 3.2 represents plots of the density of $\text{GSN}(\lambda_1, \lambda_2, \lambda_3)$. This figure further illustrates some of the simple properties of the distribution. A comparison of the density GSN(0,0,0) (Normal case) with GSN(1,0,-1) shows that the latter is also symmetric but with tails different from the normal. It appears that the distribution $\text{GSN}(\lambda_1, \lambda_2, -\lambda_1)$ can model heavier or lighter tails than the normal distribution depending on the values of λ_1. In this case, skewness is controlled by λ_2. Since this thesis is concerned with modeling skewness, further investigation of the properties of this skew symmetric model is beyond its scope.

We now investigate a new class of three-parameter generalized skew-normal distribution which does not have a link with the CSN distribution.

3.3.2 Extended two-parameter generalized skew-normal distribution

Definition 7. *A random variable $Z_{\lambda_0,\lambda_1,\lambda_2}$ is said to have an extended two-parameter generalized skew-normal distribution, if its PDF is*

$$f(z; \lambda_0, \lambda_1, \lambda_2) = k(\lambda_0, \lambda_1, \lambda_2)\phi(z)\Phi(\lambda_1 z)\Phi(\lambda_0 + \lambda_2 z), \quad z \in \mathbb{R}, \tag{3.7}$$

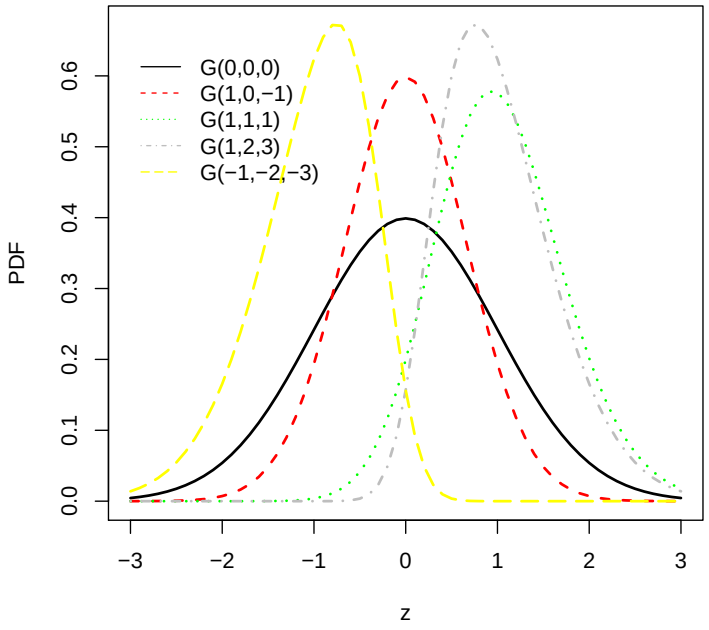

Figure 3.2: Comparison of generalized skew-normal densities

where $\lambda_0, \lambda_1, \lambda_2 \in \mathbb{R}$. λ_1 *&* λ_2 *are the skewness parameter and* λ_0 *is the shift parameter.*

Since (3.7) is a PDF, we must have

$$k(\lambda_0, \lambda_1, \lambda_2) = \frac{1}{\int_{-\infty}^{\infty} \phi(z)\Phi(\lambda_1 z)\Phi(\lambda_0 + \lambda_2 z)\, dz} = \frac{1}{E[\Phi(\lambda_1 X)\Phi(\lambda_0 + \lambda_2 X)]}, \quad (3.8)$$

where $X \sim N(0,1)$. Direct integration yields

$$\frac{1}{\Phi_2\Big(0, \frac{\lambda_0}{\sqrt{1+\lambda_2^2}}; \frac{\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda_2^2}}\Big)} = \frac{2}{\Phi_{SN}\Big(\frac{\lambda_0}{\sqrt{1+\lambda_2^2}}; 0, 1, \frac{-\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2+\lambda_2^2}}\Big)},$$

where Φ_2 is the standard bivariate normal CDF and Φ_{SN} is the standard CDF of the Azzalini (1985) SN distribution . The evaluation of Φ_{SN} can be obtained from the 'psn' function in Azzalini's SN package in R.

Thus, the extended two-parameter generalized skew-normal density in (3.7)

becomes

$$f(z;\lambda_0,\lambda_1,\lambda_2)=\frac{2}{\Phi_{SN}\Big(\frac{\lambda_0}{\sqrt{1+\lambda_2^2}};0,1,\frac{-\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2+\lambda_2^2}}\Big)}\phi(z)\Phi(\lambda_1 z)\Phi(\lambda_0+\lambda_2 z),\quad z\in\mathbb{R}, \tag{3.9}$$

and we write $Z_{\lambda_0,\lambda_1,\lambda_2}\sim \text{EGSN}(\lambda_0,\lambda_1,\lambda_2)$.

Remark 1 : For the special case $\lambda_0=0$, (3.9) becomes

$$\frac{2}{\Phi_{SN}\Big(0;0,1,\frac{-\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2+\lambda_2^2}}\Big)}\phi(z)\Phi(\lambda_1 z)\Phi(\lambda_2 z),\quad z\in\mathbb{R},$$

which is equivalent to (3.5). To see this, we note that

$$\frac{2\pi}{\cos^{-1}\Big(\frac{-\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda_2^2}}\Big)}=\frac{1}{\Phi_2\Big(0,0;\frac{\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda_2^2}}\Big)}=\frac{2}{\Phi_{SN}\Big(0;0,1,\frac{-\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2+\lambda_2^2}}\Big)}. \tag{3.10}$$

The R.H.S in (3.10) is a more general expression when the centered orthant probabilities rule is not applicable. The EGSN distribution is so named because it extends the two-parameter generalized skew-normal distribution, in the same way the ESN distribution extends the Azzalini (1985) SN distribution.

Basic properties of EGSN$(\lambda_0,\lambda_1,\lambda_2)$

Some properties of the model in (3.9) are stated below

1. EGSN$(0,\lambda_1,\lambda_2)$ = GSN(λ_1,λ_2)
2. EGSN$(\lambda_0,0,\lambda)$ = ESN(λ_0,λ)
3. EGSN$(0,0,\lambda)$ = EGSN$(0,\lambda,0)$= SN(λ)
4. EGSN$(0,0,0)$ = N(0,1)
5. EGSN$(\lambda_0,\lambda_1,\lambda_2)$ can be derived from the convolution of an independent SN random variable and a truncated normal random variable.

Moment generating function of EGSN$(\lambda_0, \lambda_1, \lambda_2)$

Theorem 3. *If $M(t; \lambda_0, \lambda_1, \lambda_2)$ is the moment generating function of $Z_{\lambda_0,\lambda_1,\lambda_2} \sim$ EGSN$(\lambda_0, \lambda_1, \lambda_2)$, then*

$$M(t; \lambda_0, \lambda_1, \lambda_2) = k(\lambda_0, \lambda_1, \lambda_2) e^{t^2/2} \Phi_2\left(\frac{\lambda_1 t}{\sqrt{1+\lambda_1^2}}, \frac{\lambda_0 + \lambda_2 t}{\sqrt{1+\lambda_2^2}}; \frac{\lambda_1 \lambda_2}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda_2^2}} \right) \tag{3.11}$$

where $k(\lambda_0, \lambda_1, \lambda_2)$ is as given in (3.9).

Proof. From (3.7), we have the MGF as

$$\begin{aligned} E(e^{tZ}) =& k(\lambda_0, \lambda_1, \lambda_2) \int_{-\infty}^{\infty} e^{tz} \phi(z) \Phi(\lambda_1 z) \Phi(\lambda_0 + \lambda_2 z)\, dz \\ =& k(\lambda_0, \lambda_1, \lambda_2) e^{t^2/2} \int_{-\infty}^{\infty} \phi(z-t) \Phi(\lambda_1 z) \Phi(\lambda_0 + \lambda_2 z)\, dz \end{aligned}$$

Put $x = z - t$. Then,

$$\begin{aligned} E(e^{tZ}) =& k(\lambda_0, \lambda_1, \lambda_2) e^{t^2/2} \int_{-\infty}^{\infty} \phi(x) \Phi(\lambda_1 x + \lambda_1 t) \Phi(\lambda_0 + \lambda_2 x + \lambda_2 t)\, dx \\ =& k(\lambda_0, \lambda_1, \lambda_2) e^{t^2/2} E(\Phi(\lambda_1 X + \lambda_1 t) \Phi(\lambda_0 + \lambda_2 X + \lambda_2 t) \\ =& k(\lambda_0, \lambda_1, \lambda_2) e^{t^2/2} P(Y_1 - \lambda_1 X < \lambda_1 t, Y_2 - \lambda_2 X < \lambda_0 + \lambda_2 t) \\ =& k(\lambda_0, \lambda_1, \lambda_2) e^{t^2/2} \Phi_2\left(\frac{\lambda_1 t}{\sqrt{1+\lambda_1^2}}, \frac{\lambda_0 + \lambda_2 t}{\sqrt{1+\lambda_2^2}}; \frac{\lambda_1 \lambda_2}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda_2^2}} \right) \end{aligned}$$

where X, Y_1, Y_2 are *iid* $N(0,1)$, and

$$P(Y_1 - \lambda_1 X < \lambda_1 t, Y_2 - \lambda_2 X < \lambda_0 + \lambda_2 t) = \Phi_2\left(\frac{\lambda_1 t}{\sqrt{1+\lambda_1^2}}, \frac{\lambda_0 + \lambda_2 t}{\sqrt{1+\lambda_2^2}}; \frac{\lambda_1 \lambda_2}{\sqrt{1+\lambda_1^2}\sqrt{1+\lambda_2^2}} \right).$$

The moments of $Z_{\lambda_0,\lambda_1,\lambda_2}$ can be obtained from (3.11). The mean and the variance of the extended two-parameter generalized skew-normal distribution is respectively,

$$\begin{aligned} E(Z_{\lambda_0,\lambda_1,\lambda_2}) =& k(\lambda_0, \lambda_1, \lambda_2) \Bigg\{ \frac{1}{\sqrt{2\pi}} \frac{\lambda_1}{\sqrt{1+\lambda_1^2}} \Phi\left(\frac{\lambda_0 \sqrt{1+\lambda_1^2}}{\sqrt{1+\lambda_1^2+\lambda_2^2}} \right) \\ &+ \frac{\lambda_2}{\sqrt{1+\lambda_2^2}} \phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_2^2}} \right) \Phi\left(\frac{-\lambda_0 \lambda_1 \lambda_2}{\sqrt{1+\lambda_2^2}\sqrt{1+\lambda_1^2+\lambda_2^2}} \right) \Bigg\}, \end{aligned} \tag{3.12}$$

$$\begin{aligned}\mathrm{Var}(Z_{\lambda_0,\lambda_1,\lambda_2}) = 1 + k(\lambda_0,\lambda_1,\lambda_2)\Bigg\{ & \frac{\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2+\lambda_2^2}}\Bigg[\frac{1}{\sqrt{2\pi}}\Big(\frac{1}{1+\lambda_1^2}\Big)\phi\Bigg(\frac{\lambda_0\sqrt{1+\lambda_1^2}}{\sqrt{1+\lambda_1^2+\lambda_2^2}}\Bigg) \\ & + \Big(\frac{1}{1+\lambda_2^2}\Big)\phi\Bigg(\frac{\lambda_0}{\sqrt{1+\lambda_2^2}}\Bigg)\phi\Bigg(\frac{-\lambda_0\lambda_1\lambda_2}{\sqrt{1+\lambda_2^2}\sqrt{1+\lambda_1^2+\lambda_2^2}}\Bigg)\Bigg] \\ & - \frac{\lambda_0\lambda_2^2}{(1+\lambda_2^2)^{3/2}}\phi\Bigg(\frac{\lambda_0}{\sqrt{1+\lambda_2^2}}\Bigg)\Phi\Bigg(\frac{-\lambda_0\lambda_1\lambda_2}{\sqrt{1+\lambda_2^2}\sqrt{1+\lambda_1^2+\lambda_2^2}}\Bigg)\Bigg\} \\ & - \Big(E(Z_{\lambda_0,\lambda_1,\lambda_2})\Big)^2.\end{aligned} \tag{3.13}$$

To fit the extended two-parameter skew-normal distribution to data, one can introduce the affine transformation $Y = \mu + \sigma Z_{\lambda_0,\lambda_1,\lambda_2} \sim \mathrm{EGSN}(\mu,\sigma^2,\lambda_0,\lambda_1,\lambda_2)$. The density becomes

$$f(y;\mu,\sigma^2,\lambda_0,\lambda_1,\lambda_2) = \frac{\frac{2}{\sigma}\phi\big(\frac{y-\mu}{\sigma}\big)\Phi\Big(\frac{\lambda_1(y-\mu)}{\sigma}\Big)\Phi\Big(\lambda_0+\lambda_2\big(\frac{y-\mu}{\sigma}\big)\Big)}{\Phi_{SN}\Big(\frac{\lambda_0}{\sqrt{1+\lambda_2^2}};0,1,\frac{-\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2+\lambda_2^2}}\Big)}. \tag{3.14}$$

The log-likelihood function in this case is

$$\begin{aligned} l(\Xi) =& n\ln 2 - \frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2} + \sum_{i=1}^{n}\ln\Phi\Big(\frac{\lambda_1(y_i-\mu)}{\sigma}\Big) \\ & + \sum_{i=1}^{n}\ln\Big[\Phi\Big(\lambda_0+\lambda_2\big(\frac{y_i-\mu}{\sigma}\big)\Big)\Big] \\ & - n\ln\Big[\Phi_{SN}\Big(\frac{\lambda_0}{\sqrt{1+\lambda_2^2}};0,1,\frac{-\lambda_1\lambda_2}{\sqrt{1+\lambda_1^2+\lambda_2^2}}\Big)\Big],\end{aligned} \tag{3.15}$$

where $\Xi = (\mu,\sigma,\lambda_0,\lambda_1,\lambda_2)$.

Since the EGSN model is an extension of the ESN model, it suffers from parameters identifiability draw-backs as well. For instance, if $\lambda_1 = \lambda_2 = 0$, the distribution becomes the normal distribution regardless of the value of λ_0. However, it is unlikely that this will be the case in practice because the two (skewness) parameters are distinct. Also, the introduction of extra parameters to a model, although leads to a more flexible model, comes at a cost of model identifiability in some cases. There may not be sufficient information in the data to identify all the parameters. The use of profile likelihood is therefore recommended to study the uncertainty in the MLEs of skew-normal models in practice.

We assess the performances of the use of skew-normal distributions to model data arising from sample selection in a simulation study. The data set is generated in a similar way as was done in the simulation study of Marchenko and Genton (2012),

but with skew-normal errors. The outcome equation is $Y_i^\star = 0.5 + 1.5x_i + \varepsilon_{1i}$, and the selection equations are $S_i^\star = 1 + x_i + \varepsilon_{2i}$ and $S_i^\star = 1 + x_i + 1.5w_i + \varepsilon_{2i}$, where $x_i \overset{iid}{\sim} N(0,1)$, $w_i \overset{iid}{\sim} N(0,1)$ and $i = 1, \ldots, N = 1000$. The use of the first selection equation ensures that all variables that predict missingness are included in the outcome equation, whereas the second selection equation has an extra predictor of missingness that is not included in the outcome model. The parameters of interest are from the outcome equation, i.e. $\beta' = (0.5, 1.5)$. The covariates x_i and w_i are independent and are also independent of the error terms ε_{1i} and ε_{2i}. The error terms are generated from bivariate skew-normal distribution with λ= 0, 0.5, 1 and 2. The covariance matrix $\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}$, where $\sigma = 1$ and the correlation $\rho = 0.5$. About 20% and 30% observations are missing when the first and the second selection equations are used respectively for data generation. Simulation results are based on 1000 replications.

Tables 3.1 and 3.2 show the finite sample performances of fitting the Azzalini (1985) SN, ESN and EGSN models to selectively reported outcomes, when the selection equation has the same parameters and an extra parameter as the outcome model respectively. We present parameters from the outcome equation only in the Tables. Parameters from the selection equations are captured by λ_0 and λ_1 when hidden truncation models are used in sample selection settings. The models' performances are similar in the two tables. The ESN model appears to outperform other models at $\lambda = 0$. The intercept of the model has less bias compared to the scenario where $\lambda \neq 0$, and this is due to the fact that the ESN model is the correct model when the underlying assumption is bivariate normal (see (3.2)). The performance of the models in identifying the intercept is poor at $\lambda = 0.5$. This has to do with the model's inability to distinguish the MLEs at that point from $\lambda = 0$, which is always a solution to the score equation. As $\lambda \to \infty$, the bias in the intercept tends to zero.

Application to the NDI scores

We fitted SN, ESN and EGSN models to the NDI scores at month 8. Table 3.3 shows the results of fitting these models. The EGSN model is constrained such that $\lambda_1 = \lambda_2$. The parameter labeled 'physio' is the Physiotherapy treatment effects. An adjustment was made for measurements at month 4 in the model, which we label 'prev'. There is a significant treatment effects according to the three model at 5% level of significance. The gender effect is not significant. A likelihood ratio test between the SN and the EGSN model gave a non-significant p-value (0.286). The

Table 3.1: Simulation results (multiplied by 10,000) using skew-distributions to model selectively reported data. Selection and Outcome equations have the same covariates.

		Bias			MSE		
		SN	ESN	EGSN	SN	ESN	EGSN
$\lambda = 0.0$	β_0	452	374	1283	2536	2280	1811
	β_1	-1475	-1471	-1478	234	283	236
	σ	889	600	338	136	110	276
$\lambda = 0.5$	β_0	2467	2055	3870	2447	3244	2811
	β_1	-1173	-1178	-1168	151	152	151
	σ	298	126	-279	58	99	177
$\lambda = 1.0$	β_0	1364	1699	2663	656	3246	1967
	β_1	-882	-886	-885	88	89	92
	σ	-130	-325	-447	53	163	396
$\lambda = 2.0$	β_0	636	-54	795	71	3315	204
	β_1	-575	-574	-575	40	40	40
	σ	-15	24	-84	20	220	42

Table 3.2: Simulation results (multiplied by 10,000) using skew-distributions to model selectively reported data. Selection equation has one more covariate that is not in Outcome equation.

		Bias			MSE		
		SN	ESN	EGSN	SN	ESN	EGSN
$\lambda = 0.0$	β_0	1272	641	1349	2976	2311	1881
	β_1	-691	-685	-686	64	64	64
	σ	1115	759	493	186	126	178
$\lambda = 0.5$	β_0	3340	2803	3423	3352	3384	3458
	β_1	-579	-578	-580	47	47	48
	σ	411	179	402	66	85	134
$\lambda = 1.0$	β_0	1499	1646	1516	908	3445	1095
	β_1	-439	-441	-438	30	30	31
	σ	-87	-279	-40	56	158	164
$\lambda = 2.0$	β_0	529	192	494	63	3317	136
	β_1	-288	-287	-286	15	15	16
	σ	63	-1	101	25	221	65

SN model can therefore be used to describe this data.

Table 3.3: Fit of Azzalini (1985) model, ESN and EGSN model to complete case NDI scores at 8 months. λ_1 and λ_2 are constrained to be equal in the EGSN model.

	SN			ESN			EGSN		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
int	5.593	0.940	0.000	7.915	5.671	0.164	5.388	0.679	0.000
age	0.083	0.024	0.001	0.082	0.024	0.001	0.087	0.017	0.000
sex(f)	0.569	0.646	0.379	0.623	0.642	0.333	0.636	0.455	0.163
prev	0.667	0.039	0.000	0.668	0.040	0.000	0.665	0.027	0.000
physio	1.479	0.617	0.017	1.820	0.613	0.003	1.477	0.434	0.001
σ	8.796	0.584	0.000	9.429	1.497	0.000	8.627	0.418	0.000
λ_1	-1.295	0.273	0.000	-1.442	0.410	0.001	-1.234	0.198	0.000
λ_0	-	-	-	-0.646	1.374	0.638	5.195	4.198	0.217
Loglik	-1596.86			-1596.32			-1596.29		

3.4 Modeling bounded scores with truncated skew-normal distribution

The results of the skew-normal models fitted to the NDI scores at month 8 (see Table 3.3), did not take into account the lower and upper bounds of the data. In practice, a properly fitted distribution is expected to cover the range of values over which the model variable could theoretically extend. If a fitted distribution extends beyond the range of plausible values, then the model will produce unrealistic values at the extreme tails of the distribution.

All scores in the NDI data belong to the interval [0, 50], and skewness is apparent in the data. There are many strategies available in the literature to model such skew and bounded outcome. One strategy is to use transformation (e.g. logistic transformation) and then model the transformed data using a skew distribution. However, as we remark in chapter 1, transformation of the data may not remove the non-linear dependence of the transformed scores on covariates. In cases where the truncation bounds are known, it may be natural to model skew bounded scores using truncated distributions.

3.4.1 Truncated distributions

Suppose we have a continuous distribution with PDF and CDF specified as $g(.)$ and $G(.)$, respectively. Let Y be a random variable representing the truncated version of this distribution over the interval [a, b], where $-\infty < a < b < \infty$. The PDF and CDF of Y are given respectively by

$$f_Y(y) = \begin{cases} \frac{g(y)}{G(b)-G(a)} & \text{if } a \le y \le b, \\ 0 & \text{otherwise} \end{cases}$$

and

$$F_Y(y) = \frac{G(\max(\min(y,b),a)) - G(a)}{G(b) - G(a)}.$$

Expression for the mean, variance, quantile function and generation of random numbers from truncated distributions can be found in Nadarajah and Kotz (2006). In particular, the truncated extended skew-normal (TESN) distribution has a standard PDF given by

$$f_{TESN}(\lambda_0, \lambda_1, a, b) = c(\lambda_0, \lambda_1, a, b)\phi(z)\Phi(\lambda_0 + \lambda_1 z), \tag{3.16}$$

where

$$c(\lambda_0, \lambda_1, a, b) = \frac{1}{\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)\Big[\Phi_{ESN}(b; \lambda_0, \lambda_1) - \Phi_{ESN}(a; \lambda_0, \lambda_1)\Big]}.$$

The corresponding CDF is given by

$$F_{TESN}(\lambda_0, \lambda_1, a, b) = \begin{cases} 0 & \text{if } z < a, \\ \frac{\Phi_{ESN}(z,\lambda_0,\lambda_1) - \Phi_{ESN}(a,\lambda_0,\lambda_1)}{\Phi_{ESN}(b,\lambda_0,\lambda_1) - \Phi_{ESN}(a,\lambda_0,\lambda_1)} & \text{if } a \le z < b, \\ 1 & \text{if } z \ge b. \end{cases}$$

The expression for Φ_{ESN} is not readily available in statistical software, but can be easily computed from the CDF of multivariate normal distribution. If $Y \sim \text{ESN}(\mu, \sigma^2, \lambda_0, \lambda_1)$, then it has a closed skew-normal form $CSN_{1,1}(\mu, \sigma^2, \lambda_1/\sigma, -\lambda_0, 1)$. The corresponding CDF can be computed using equation (2.12), and we have

$$\frac{1}{\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)}\Phi_2\left(\begin{pmatrix} y \\ 0 \end{pmatrix}; \begin{pmatrix} \mu \\ -\lambda_0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & -\lambda_1\sigma \\ -\lambda_1\sigma & 1+\lambda_1^2 \end{pmatrix}\right).$$

3.4.2 Truncated skew-normal distribution and the NDI scores

Truncated skew-normal (TSN) distribution has been discussed in the literature (see Kim (2004), Jamalizadeh et al. (2009) and Flecher et al. (2010)). The model is a realistic model for the NDI scores at month 8 since skewness is apparent in the data and the floor and ceiling effects in the data can be adjusted for. If $\lambda_0 =$

0 in (3.16), the TSN model is recovered. Table 3.4 shows the results of fitting regression models with truncated normal and TSN error distributions to the NDI scores. The truncation points are taken into account in the model with the lower and upper bounds taken to be 0 and 50 respectively. The truncated normal error is fitted for comparison purposes only as the data is clearly skew (likelihood ratio test gave p-value <0.0001). A comparison of TSN model (Table 3.4) and the SN model (Table 3.3) using LRT shows that the TSN model fits better. Although the statistical significance of the parameters in the SN and TSN models are the same, the parameters in the TSN model are consistently larger in magnitude. This is due to the restricted range over which the parameters are maximized.

Table 3.4: Fit of truncated normal (TN) and truncated skew-normal (TSN) models to complete case NDI scores at 8 months.

	TN			TSN		
	Estimate	S.E.	p-value	Estimate	S.E	p-value
int	-12.561	2.620	0.000	-0.287	1.830	0.876
age	0.171	0.047	0.000	0.146	0.041	0.001
sex(f)	1.303	1.303	0.318	0.756	1.096	0.491
prev	1.052	0.091	0.000	1.044	0.077	0.000
physio	2.643	1.230	0.032	2.724	1.050	0.010
σ	9.441	0.587	0.000	19.419	3.345	0.000
λ_1	-	-	-	-3.202	0.749	0.000
Loglik	-1496.28			-1483.45		

3.5 Summary

We have written down two types of three-parameter generalized skew-normal distributions, which are extensions of the two-parameter generalized skew-normal distribution of Jamalizadeh et al. (2008). The first of these models is a special case of the CSN distribution, which can model skewness and tail-weight simultaneously. Since the focus of this thesis is on modeling skewness, we have not studied statistical properties of the model, and in particular the characterization of the tail-weight. The second distribution (EGSN) does not have direct link with the CSN distribution. This model is the basis of our model in chapter 4.

Finite sample performances of skew-normal distributions in modeling data arising from sample selection were examined in a simulation study. The link between sample selection, hidden truncation and skew distributions implies that skew

distributions can be used to model complete cases in selectively reported outcomes. Although the use of skew distributions are justifiable mathematically, parameters effect may not be completely accounted for in the model. For instance, data arising from sample selection with underlying bivariate normal assumption mathematically results in equation (3.2). The use of ESN distribution to model the data, from hidden truncation perspectives, implies that the function $\gamma' x/\sqrt{1-\rho^2}$ is modeled as a single parameter λ_0. In principle, $\gamma' x$ carries covariate information, which cannot be fully adjusted for in λ_0. It is therefore necessary to take into account the data generation process before proposing models, rather than using models based on their mathematical links. We also examined MAR scenarios, where $\rho = 0$ (not shown here.) As expected, the three models gave better fit with almost no bias when compared with the $\rho = 0.5$ cases given in Tables 3.1 and 3.2.

Since the data is on a finite range, the use of TSN model was proposed. The model gave a better fit and its predictive capability is superior to it non-truncated counterparts. The interpretation of the parameter is on the original scale unlike what we might have obtained with data transformation.

The use of skew-normal distributions for modeling data arising from sample selection is not recommended in practice. This can leads to inflated type 1 error, where parameters in the model becomes significant, when in fact they are not. The treatment effect is significant in all the complete case models we considered in this chapter. This is shown not to be true when a full sample selection model (i.e. the missingness process is included in the model) is used, as we show in next chapter.

Chapter 4

A Sample Selection Model With Skew-Normal Distribution

In chapter 3 we mention that modeling data sets arising from sample selection using skew distributions amount to complete case analysis. Although parameter estimates may be unbiased, there could be inflated type-1 error in statistical significance tests. We show in this chapter that the additional information about observability or non-observability of the data included in classical sample selection density can correct for this error, and aid model identifiability. In particular, we develop a sample selection model with underlying skew-normal distribution. A link is established between the continuous component of our model log-likelihood function and the extended two-parameter generalized skew-normal distribution introduced in chapter 3. This link is used to derive the expected value of the model, which extends Heckman's two-step method. Finite sample performance of the maximum likelihood estimator of the model is studied via Monte Carlo simulation. The model is applied to the NDI scores at month 8 and month 12. The application of the model to scores at month 12 is to emphasis the influence of conditional normality in sample selection models. We discuss computational and identification issues, and give directions for possible extensions of the model.

4.1 Sample selection models

Recall the regression models given in section 3.1, that is

$$Y_i^\star = \beta' x_i + \sigma\varepsilon_{1i}, \quad i = 1, \ldots, N, \tag{4.1}$$

as regression model of interest, and selection mechanism given as

$$S_i^\star = \gamma' x_i + \varepsilon_{2i}, \quad i = 1, \ldots, N, \tag{4.2}$$

where $\beta, \gamma, x_i, Y_i^\star, S_i^\star, Y, S$ and n are as defined in section 3.1. Under the bivariate normal assumption of the error terms ε_{1i} and ε_{2i}, the conditional density $f(y|x, S = 1; \Theta)$ (where $\Theta = (\beta, \sigma, \gamma, \rho)$) is given by equation (3.2). This equation is not the full sample selection density. The density of the sample selection model is composed of a continuous component corresponding to the conditional density $f(y|x, S = 1; \Theta)$ and a discrete component given by $P(S = 1|x)$. The marginal distribution of the selection equation determines the nature of the model to be fitted to the discrete component. In Copas and Li (1997) (and Heckman (1976)), a probit model $P(S = s) = \{\Phi(\gamma' x)\}^s \{1 - \Phi(\gamma' x)\}^{1-s}$ was used. The log-likelihood function is therefore

$$l(\Theta) = \sum_{i=1}^{n} S_i \Big(\ln f(y_i|x_i, S_i = 1; \Theta) \Big) + \sum_{i=1}^{n} S_i (\ln \Phi(\gamma' x_i)) + \sum_{i=1}^{n} (1 - S_i) \ln \Phi(-\gamma' x_i). \tag{4.3}$$

The maximum likelihood estimation based on (4.3) is not robust to deviations from the normality assumption. This prompted Heckman (1979) to develop the two-step estimator (TS). The TS estimator is derived from the conditional expectation of the observed data, and is given by

$$E(Y|x, S^\star > 0) = \beta' x + \sigma \rho \Lambda(\gamma' x), \tag{4.4}$$

where Λ is the inverse Mills ratio. This model is equivalent to equation (3.4) when $\gamma' x = \lambda_0 \sqrt{1 - \rho^2}$ and $\rho = \lambda_1 / \sqrt{1 + \lambda_1^2}$. To use (4.4) in practice, a standard probit model for S provides an estimate of $\hat{\gamma}$. The quantity $\Lambda(\hat{\gamma}' x)$ is then taken as an additional covariate in equation (4.4), and the least squares coefficient of $\Lambda(\hat{\gamma}' x)$ gives an estimate of $\sigma\rho$.

The TS method is moment based and does not require distributional assumption for the error terms in the second-step OLS procedure to obtain consistent estimator. However, when the outcome and the selection equations contain the same covariates, the method has been shown to perform poorly due to multicollinearity (see Puhani (2000)). This is because the inverse Mills ratio is nearly linear over a wide range of its support. To avoid this problem in practice, an *exclusion restriction*, where at least one extra variable is a good predictor of non-response is included in the selection equation and excluded from the primary regression.

The conditional variance of the observed data can be derived using the link

between the ESN density and equation (3.2). This gives

$$\text{var}(y|x, S^\star > 0) = \sigma^2[1 - \rho^2\Lambda(\gamma' x)\{\gamma' x + \Lambda(\gamma' x\}]. \tag{4.5}$$

To obtain the estimates of ρ and σ in (4.4), the average value of the right-hand side of (4.5) is equated to the observed residual of the second-step regression. Note that the mean depends linearly on ρ but the variance does not. Thus, most of the parameters of interest may be sensitive to small changes in the value of ρ.

As noted in chapter 3, the continuous component of the sample selection density (equation (3.2)) is essentially an ESN density (equation (2.8)). The ESN distribution is not identifiable when $\lambda = 0$ ($\rho = 0$ in the case of (3.2)) but the model becomes identifiable in the sample selection framework due to the additional information from the selection process which is introduced through a probit model. The price to pay for the identifiability is possibility of model misspecification. Although sensitivity analysis on the model parameters is justifiable, the use of range of plausible parametric representations, especially those having the normal distribution as special case, is preferred. In the following section, we develop a sample selection model with an underlying skew-normal error distribution.

4.2 Selection Skew-normal model (SSNM)

In this section, we relax the assumption of bivariate normality of the Heckman (1976) model such that the underlying error distribution is bivariate skew-normal. We show that the continuous component of our model log-likelihood function can be derived using conditioning approach of equation 3.1 or the hidden truncation of Arnold and Beaver (2002), and that the methods are equivalent. This link is used to derive a Heckman-type two-step estimation method under the skew-normal distribution.

4.2.1 Conditioning in bivariate skew-normal distribution to formulate SSNM model

The continuous component of the sample selection density given by (3.2) was derived using the conditional distribution properties of a bivariate normal distribution. Suppose we relax the assumption of bivariate normality given in section 3.1 such that the underlying error distribution is bivariate skew-normal. i.e

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim SN_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \right\},$$

where λ_1 and λ_2 are the skewness parameters for $Y_i^\star$ and $S_i^\star$ respectively. Then $f(y|x, S = 1; \Xi)$ (where $\Xi = \beta, \sigma, \gamma, \rho, \lambda_1, \lambda_2$) is still defined as equation (3.1). To determine the expression $P(S^\star > 0|y, x)$ in equation (3.1), it is easier to write the joint distribution in the CSN form, that is

$$\begin{pmatrix} Y \\ S \end{pmatrix} \sim CSN_{2,1}\Big\{\boldsymbol{\mu} = (\beta' x, \gamma' x), \Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}, D = (\lambda_1/\sigma, \lambda_2), \nu = 0, \Delta = 1\Big\}. \tag{4.6}$$

The distribution of $S|Y$, using the conditional distribution property of CSN (see equation (2.14)), is

$$S|Y \sim CSN_{1,1}\Big\{\gamma' x + \rho\Big(\frac{y - \beta' x}{\sigma}\Big), 1 - \rho^2, \lambda_2, -(\lambda_1 + \lambda_2)\Big(\frac{y - \beta' x}{\sigma}\Big), 1\Big\},$$

and $P(S^\star > 0|Y)$ is an ESN lower tail probability written as

$$\Phi_{ESN}\Big\{\gamma' x + \rho\Big(\frac{y - \beta' x}{\sigma}\Big); 0, 1 - \rho^2, \frac{-\lambda_2}{\sqrt{1 - \rho^2}}, -(\lambda_1 + \lambda_2)\Big(\frac{y - \beta' x}{\sigma}\Big)\Big\}. \tag{4.7}$$

To determine the expression $P(S^\star > 0)$ in equation (3.1) we need to extract its marginal distribution from the bivariate process. Using the property of marginalization of CSN (see equation (2.13)), we have

$$P(S^\star > 0) = \Phi_{SN}\left(\gamma' x; 0, 1, \frac{-(\lambda_2 + \lambda_1\rho)}{\sqrt{1 + \lambda_1^2 - \lambda_1^2\rho^2}}\right), \tag{4.8}$$

where Φ_{SN} denotes the CDF of a skew-normal random variable. The marginal distribution of the outcome equation is

$$Y \sim CSN_{1,1}\Big\{\beta' x, \sigma^2, \Big(\frac{\lambda_1 + \lambda_2\rho}{\sigma}\Big), 0, (1 + \lambda_2^2 - \lambda_2^2\rho^2)\Big\},$$

and the corresponding PDF is

$$f(y) = \frac{2}{\sigma}\phi\left(\frac{y - \beta' x}{\sigma}\right)\Phi\Big\{\Big(\frac{\lambda_1 + \lambda_2\rho}{\sqrt{1 + \lambda_2^2 - \lambda_2^2\rho^2}}\Big)\Big(\frac{y - \beta' x}{\sigma}\Big)\Big\}. \tag{4.9}$$

Substituting (4.7), (4.8) and (4.9) into the general sample selection equation

(3.1) we have $f(y|x, S = 1; \Xi)$ given by

$$\frac{f(y)\Phi_{ESN}\left\{\gamma'x + \rho\left(\frac{y-\beta'x}{\sigma}\right); 0, 1-\rho^2, \frac{-\lambda_2}{\sqrt{1-\rho^2}}, -(\lambda_1+\lambda_2)\left(\frac{y-\beta'x}{\sigma}\right)\right\}}{\Phi_{SN}\left(\gamma'x; 0, 1, \frac{-(\lambda_2+\lambda_1\rho)}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}}\right)}. \tag{4.10}$$

If λ_1 and λ_2 are set equal to zero in (4.10), Copas and Li (1997) model given by (3.2) is recovered.

From now on, we shall restrict attention to a special case of the model given in (4.10). Suppose only λ_2 is set equal to zero, (i.e. selection random variable is normal) we get a simpler model:

$$f(y|x, S = 1; \Omega) = \frac{\frac{2}{\sigma}\phi\left(\frac{y-\beta'x}{\sigma}\right)\Phi\left(\frac{\lambda_1(y-\beta'x)}{\sigma}\right)\Phi\left(\frac{\gamma'x+\rho\left(\frac{y-\beta'x}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)}{\Phi_{SN}\left(\gamma'x; 0, 1, \frac{-\lambda_1\rho}{\sqrt{1+\lambda_1^2-\lambda_1^2\rho^2}}\right)}, \tag{4.11}$$

where $\Omega = (\beta, \sigma, \gamma, \rho, \lambda_1)$. This situation is possible in practice when the underlying mechanism governing selection is not skewed before entering the joint process.

Equation (4.11) is the basis of the *extended two-parameter generalized skew-normal* (EGSN) density introduced in equation (3.9). This model is the continuous component of the sample selection density for underlying bivariate skew-normal error distribution. The model can be readily derived using the hidden truncation approach as we show below.

4.2.2 Hidden truncation formulation of SSNM model

Suppose $Z \sim SN(0, 1, \lambda_1)$ and $S \sim N(0, 1)$, with Z & S independent. Then (3.3) becomes

$$f(z|\lambda_0 + \lambda Z > S) = \frac{2\phi(z)\Phi(\lambda_1 z)\Phi(\lambda_0 + \lambda z)}{P(\lambda_0 + \lambda Z > S)}. \tag{4.12}$$

The determination of the normalizing constant $P(\lambda_0 + \lambda Z > S)$, requires the distribution of $S - \lambda Z$:

$$(S - \lambda Z) \sim \mathrm{SN}\left(0, (1+\lambda^2), \frac{-\lambda_1\lambda}{\sqrt{1+\lambda_1^2+\lambda^2}}\right). \tag{4.13}$$

Equation (4.13) was derived using the scalar multiplication and additive properties of the skew-normal distribution. Details of this can be found in equation

(2.15). So,

$$\mathrm{P}(S-\lambda Z<\lambda_0)=\Phi_{SN}\bigg(\frac{\lambda_0}{\sqrt{1+\lambda^2}};0,1,\frac{-\lambda_1\lambda}{\sqrt{1+\lambda_1^2+\lambda^2}}\bigg).$$

Equation 4.12 can now be written as

$$f(z|\lambda_0+\lambda Z>S)=\frac{2\phi(z)\Phi(\lambda_1 z)\Phi(\lambda_0+\lambda z)}{\Phi_{SN}\bigg(\frac{\lambda_0}{\sqrt{1+\lambda^2}};0,1,\frac{-\lambda_1\lambda}{\sqrt{1+\lambda_1^2+\lambda^2}}\bigg)},$$

which is equivalent to (3.14) when we make use of the transformation $Y=\mu+\sigma Z$.

Now that we have established the equivalence of the two routes of determining the continuous component of SSNM density, the discrete component can be determined by the marginal distribution of the selection equation. In this case, we have a binary regression model with the skew-normal link.

The conditional expectation and variance of the observed data can be derived from equations (3.12) and (3.13). In particular, the mean ($E(Y|x,S^\star>0)$) is given by

$$\begin{aligned}\beta'x+\sigma\Bigg[\bigg(\frac{2}{\Phi_{SN}\bigg(\gamma'x;0,1,\frac{-\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\bigg)}\bigg)\Bigg\{\frac{1}{\sqrt{2\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}\Phi\bigg(\frac{\gamma'x\sqrt{1+\lambda^2}}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\bigg)\\ +\rho\phi(\gamma'x)\Phi\bigg(\frac{-\gamma'x\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\bigg)\Bigg\}\Bigg].\end{aligned} \tag{4.14}$$

When $\lambda=0$ in equation (4.14), we have the Heckman two-step model given in equation (4.4). To visualize the impact of using selection-normal model when the correct model is the one given by equation (4.14), we plot the second component of the expectation ($E(Y|x,S^\star>0)$- $\beta'x$) as a function of $\gamma'x$, the mean of the selection variable. We take $\rho=0.5$ and 0.9 for values of λ= 0, 1, 2 and 5. It should be noted that $\lambda=0$ corresponds to the inverse Mills ratio correction in (4.4). The standard deviation, σ, simply scales the correction factor and ρ is the correlation between the outcome and the selection process.

It can be seen from Figure 4.1 ($\rho=0.5$) that especially for positive values of the selection linear predictor $\gamma'x$, the conditional expectation will be underestimated under the usual selection-normal model. This underestimation increases as the skewness increases. However, for negative values of $\gamma'x$, the underestimation of the

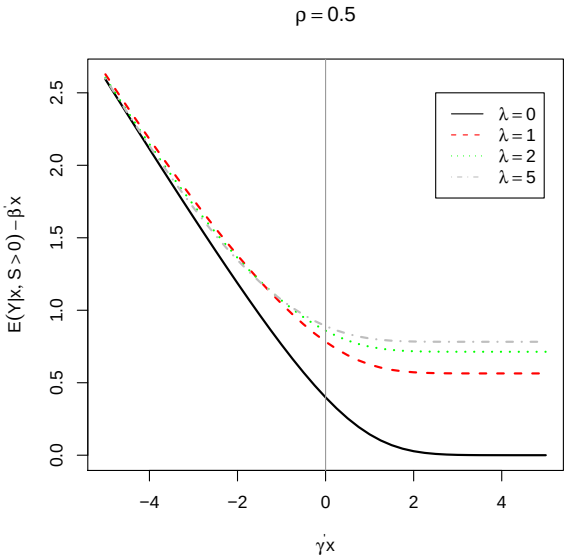

Figure 4.1: Plot of correction factor for different values of skewness parameter with $\lambda = 0$ corresponding to the normal case.

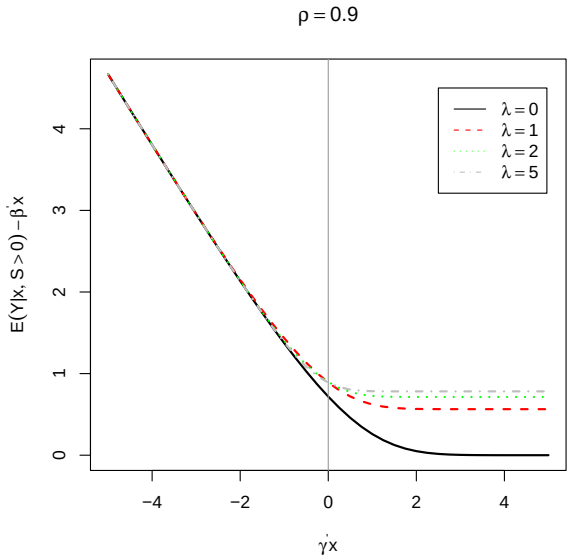

Figure 4.2: Plot of correction factor for different values of skewness parameter with $\lambda = 0$ corresponding to the normal case.

conditional expectation by the selection-normal model compared to selection skew-normal model decreases and the difference dies out as $\gamma' x$ becomes more negative and missingness increases. This observation is also true for $\rho = 0.9$, as the figures are similar (see Figure 4.2).

Sometimes, the marginal effect of the covariates (x_i) on the outcome Y_i in the observed sample may be of interest. For the Heckman two-step model, the effect consists of two components- the direct effect of the covariates on the mean of Y_i which is captured by β and the indirect effect of the covariates in the selection equation. For Heckman two-step model (equation (4.4)), the marginal effect is given by

$$\frac{\partial}{\partial x_i} E(Y|x, S^\star > 0) = \beta_i' - \rho\sigma\gamma_i' \left\{ \gamma' x \frac{\phi(\gamma' x)}{\Phi(\gamma' x)} + \left(\frac{\phi(\gamma' x)}{\Phi(\gamma' x)} \right)^2 \right\}. \tag{4.15}$$

Using similar argument, the marginal effect $(\frac{\partial}{\partial x_i} E(Y|x, S^\star > 0))$ correspond-

ing to equation (4.14) can be written as

$$\beta_i' - \sigma\gamma_i' \Bigg[\left(\frac{2}{\Phi_{SN}\left(\gamma' x; 0, 1, \frac{-\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\right)} \right) \Bigg\{ \rho(\gamma' x)\phi(\gamma' x)\Phi\left(\frac{-\gamma' x \lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}} \right)$$
$$+ \frac{2\rho\left(\phi(\gamma' x)\right)^2 \left(\Phi\left(\frac{-\gamma' x\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\right)\right)^2}{\Phi_{SN}\left(\gamma' x; 0, 1, \frac{-\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\right)} - \frac{1}{\sqrt{2\pi}} \frac{\lambda}{\sqrt{1+\lambda^2-\lambda^2\rho^2}} \phi\left(\frac{\gamma' x\sqrt{1+\lambda^2}}{\sqrt{1+\lambda^2-\lambda^2\rho^2}} \right)$$
$$+ \frac{\lambda\rho^2\phi(\gamma' x)}{\sqrt{1+\lambda^2-\lambda^2\rho^2}} \phi\left(\frac{-\gamma' x\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}} \right)$$
$$+ \frac{1}{\sqrt{2\pi}} \frac{\lambda}{\sqrt{1+\lambda^2}} \Phi\left(\frac{\gamma' x\sqrt{1+\lambda^2}}{\sqrt{1+\lambda^2-\lambda^2\rho^2}} \right) \frac{2\phi(\gamma' x)\Phi\left(\frac{-\gamma' x\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\right)}{\Phi_{SN}\left(\gamma' x; 0, 1, \frac{-\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\right)} \Bigg\} \Bigg]. \tag{4.16}$$

Equation (4.16) reduces to equation (4.15) when $\lambda = 0$.

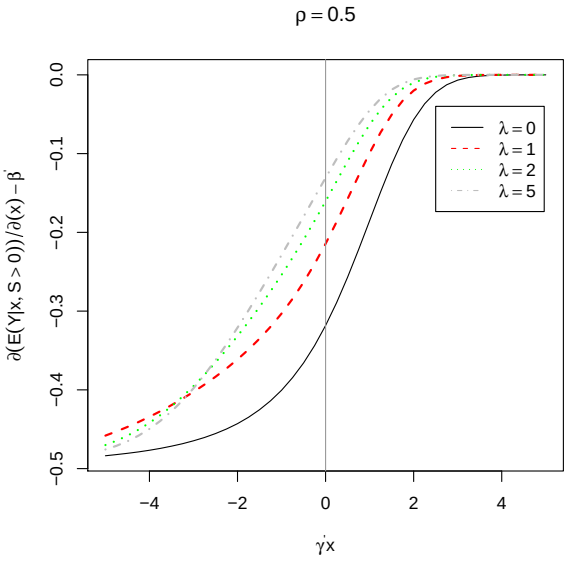

Figure 4.3: Plot of marginal effect for different values of skewness parameter with $\lambda = 0$ corresponding to the normal case.

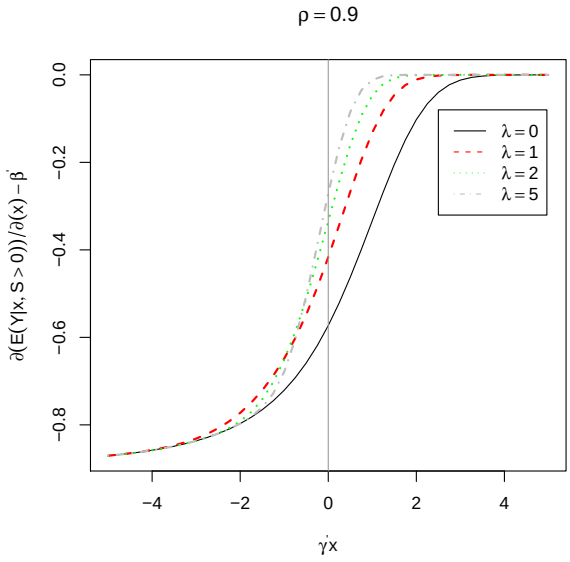

Figure 4.4: Plot of marginal effect for different values of skewness parameter with $\lambda = 0$ corresponding to the normal case.

From Figures 4.3 and 4.4, the conditional marginal effect of covariates x_i on the outcome Y will be underestimated by the selection-normal model for positive

values of $\gamma' x$ between (roughly) -4 and 4. When $|\gamma' x|$ exceeds 4, this effect dies out since the correction factor becomes zero for all the values of λ (including $\lambda = 0$).

The complete density of the selection skew-normal model, like the selection normal model, is comprised of a continuous component given by (4.11) and a discrete component for $P(S = 1|x)$. As stated earlier, the marginal distribution of the selection process determines the nature of the model to be fitted for the binary variable which in this case is given by

$$P(S = s) = \{\Phi_{SN}(\gamma' x; 0, 1, \lambda^\star)\}^s \{1 - \Phi_{SN}(\gamma' x; 0, 1, \lambda^\star)\}^{1-s},$$

where $\lambda^\star = -\lambda\rho/\sqrt{1 + \lambda^2 - \lambda^2\rho^2}$. This is a binary regression model with a skew-normal link. The log-likelihood function is therefore

$$\begin{aligned} l(\Omega) = &\sum_{i=1}^{n} S_i \Big(\ln f(y_i|x_i, S_i = 1) \Big) + \sum_{i=1}^{n} S_i \Big(\ln \Phi_{SN}(\gamma' x_i; 0, 1, \lambda^\star) \Big) \\ &+ \sum_{i=1}^{n} (1 - S_i) \ln \Phi_{SN}(-\gamma' x; 0, 1, -\lambda^\star). \end{aligned} \tag{4.17}$$

4.2.3 Monte Carlo Simulation

In this section we study finite sample properties of our selection skew-normal model (SSNM). We compare its performance with selection normal model (Heckman, 1976) SNM, and the Heckman's two-step method TS. The data generation is essentially the same as the simulation scenarios given in chapter 3. We refer to the selection equation $S_i^\star = 1 + x_i + 1.5w_i + \varepsilon_{2i}$ as scenario with exclusion restriction and $S_i^\star = 1 + x_i + \varepsilon_{2i}$ is without the exclusion restriction. The advantage of the exclusion restriction has been discussed in section 4.1.

The results of the simulation in the presence of exclusion restriction are presented in Table 4.1. Even under the normality assumption (i.e. $\lambda = 0$) the performance of SSNM is comparable to SNM and TS. For instance, SNM and TS showed slightly less bias in the estimation of the intercept of the outcome equation than SSNM. However, this advantage is counter-balanced when the intercept of the selection equation is considered since it has less bias than SNM and TS. In terms of MSE, SNM and TS are more efficient. Other parameters are comparable across the three models. In effect, SNM and TS do not appear to show emphatic superior advantage overall even with underlying normal assumption.

As the degree of skewness increases, the SSNM model gets better in precision of estimating the intercept of the selection and the outcome equations whereas SNM

and TS get worse. When $\lambda = 5$ (which is almost a folded normal), the SNM and TS break down. However, SSNM performs well but at a cost of non-convergence for some of the samples (in this case, 828 out of 1000 samples converged).

The results of the simulation in the absence of exclusion restriction are presented in Table 4.2. When the underlying process is normal, the intercept has a lower bias than SNM but higher than TS. For regression parameters of interest, the three models are comparable. Similar to what we observed under exclusion restriction, the SSNM model appears useful even when the underlying process is normal. When λ increases, the performance of SSNM gets better both in bias and MSE. There were severe identifiability problems with SSNM model when $\lambda = 5$ as about 300 samples out of 1000 produced errors in the optimization algorithm. This may be due to the fact that $\lambda = 5$ is close to the half-normal distribution.

In addition, the SSNM estimates are better than the SNM and TS models for σ and ρ when $\lambda \geq 1$ both in the presence and absence of the exclusion restriction. Since, the variance σ describes the variability of the probability distribution of the outcomes Y_i, correct prediction intervals of new observations will be obtained under SSNM model. Further, in applied settings (similar to the MINT Trials data we describe next), interest may be on patients who do not return their questionnaire. This requires a correct model for the selection process. The SSNM gave consistently smaller bias and MSE as compared to SNM and TS models for the selection equation when $\lambda \geq 1$ (Tables 4.1 and 4.2). The bias in the parameter estimates of the selection equation when SSNM model is used is smaller even under normality assumption, with or without the exclusion restriction.

We also considered the effect of varying the underlying correlation in the presence of exclusion restriction for $\lambda = 1$ and 2. The results (see Tables 1 and 2 in Appendix A) are similar to the ones for $\rho = 0.5$.

Application of selection skew-normal model to the NDI scores

Vernon (2009) recommended that patients with only 2 missed items should be considered complete, with mean imputation used for adjustment. We follow this recommendation and any patient with 3 or more missing items are considered as unit missing. In effect, we have only unit non-response left in the data set. In what follows, we will identify predictors of dropout at each measurement occasion before fitting the SSNM model to the scores at month 8 and 12.

Table 4.1: Simulation results (multiplied by 10,000) in the presence of exclusion restriction.

		Bias			MSE		
		SSNM	SNM	TS	SSNM	SNM	TS
$\lambda = 0.0$	β_0	16	-1	2	108	24	27
	β_1	-3	-3	-5	19	19	19
	γ_0	61	67	73	74	50	51
	γ_1	40	52	59	60	59	60
	γ_2	80	98	106	94	93	94
	σ	28	-9	-7	17	9	9
	ρ	-7	-6	-21	84	84	113
	λ	-27	-	-	175	-	-
$\lambda = 0.5$	β_0	2071	3564	3564	1379	1289	1291
	β_1	1	2	2	16	16	16
	γ_0	1786	2091	2101	517	507	514
	γ_1	203	259	269	74	75	78
	γ_2	314	398	409	126	125	130
	σ	-444	-654	-652	65	50	50
	ρ	-173	-243	-248	104	102	129
	λ	-30	-	-	1267	-	-
$\lambda = 1.0$	β_0	445	5620	5624	361	3173	3178
	β_1	4	10	7	12	12	12
	γ_0	401	3516	3529	282	1319	1330
	γ_1	108	533	547	73	98	102
	γ_2	201	835	860	138	192	199
	σ	-110	-1697	-1696	67	293	293
	ρ	-72	-636	-658	133	155	181
	λ	-501	-	-	1471	-	-
$\lambda = 2.0$	β_0	13	7088	7098	36	5034	5049
	β_1	7	20	14	8	9	9
	γ_0	149	4706	4728	302	2310	2333
	γ_1	86	850	877	88	151	157
	γ_2	140	1275	1324	171	304	317
	σ	-6	-2879	-2881	22	833	834
	ρ	-65	-1087	-1145	170	250	285
	λ	311	-	-	993	-	-

Table 4.2: Simulation results (multiplied by 10,000) in the absence of exclusion restriction.

		Bias			MSE		
		SSNM	SNM	TS	SSNM	SNM	TS
$\lambda = 0.0$	β_0	143	154	49	607	84	124
	β_1	-123	-121	36	62	62	89
	γ_0	-2	66	66	167	38	38
	γ_1	29	100	101	55	52	52
	σ	228	-18	59	69	12	23
	ρ	-359	-427	-237	474	452	651
	λ	18	-	-	1139	-	-
$\lambda = 0.5$	β_0	2912	3675	3593	1334	1411	1372
	β_1	-108	-88	-20	50	48	63
	γ_0	1646	2036	2038	463	461	462
	γ_1	157	217	220	60	58	58
	σ	-406	-642	-586	59	49	50
	ρ	-654	-648	-440	604	544	683
	λ	-3782	-	-	2527	-	-
$\lambda = 1.0$	β_0	640	5580	5604	381	3151	3187
	β_1	-76	48	25	37	36	42
	γ_0	759	5261	5340	527	2841	2926
	γ_1	91	434	490	73	84	88
	σ	-138	-1637	-1628	67	276	276
	ρ	-669	-604	-733	768	548	721
	λ	-761	-	-	1512	-	-
$\lambda = 2.0$	β_0	36	6812	7085	45	4681	5051
	β_1	-17	304	37	23	49	29
	γ_0	333	4451	4677	884	2052	2251
	γ_1	-47	507	865	121	114	141
	σ	33	-2708	-2827	19	741	805
	ρ	-556	100	-1245	879	869	864
	λ	165	-	-	916	-	-

Use of Probit model to identify predictors of dropout

In any model involving missing data, it is important to include covariates that are predictors of dropout in the model. For the NDI scores, we use probit regression model to identify predictors of dropout. Binary response variables were constructed with value 1 if patient drops out by months 4, 8 or 12 and 0 otherwise. The first step was to consider if the baseline measurements could influence dropout. We then consider whether any pre-randomization variables give any further improvement. The two treatments under consideration were also included with the reinforcement of advice used as the reference category. Monotone pattern of missing data is considered in order to incorporate measurement at previous occasion into the model. This results in 502, 479 and 426 observations included at months 4, 8, and 12 respectively.

The results of these models are presented in Table 4.3. Measurements at baseline, months 4 and 8 are labeled 'base', 'm4' and 'm8' respectively. We focus on the missingness model at months 8 for the moment, which shows that age and sex of the patients are good predictors of missingness. The model showed that females are more likely to drop out than males.

Table 4.3: Probit model for dropout at 4, 8 and 12 months using Vernon scores.

	Missing at 4 months			Missing at 8 months			Missing at 12 months		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
int	1.293	0.266	0.000	0.888	0.284	0.002	0.816	0.256	0.001
physio	-0.233	0.181	0.198	0.185	0.194	0.340	0.291	0.167	0.082
sex(f)	0.048	0.184	0.795	0.539	0.189	0.005	0.292	0.170	0.086
age	0.006	0.007	0.381	0.029	0.008	0.000	0.011	0.007	0.109
base	0.010	0.011	0.377	-0.010	0.015	0.510	-0.032	0.014	0.018
m4				0.025	0.015	0.100	0.016	0.015	0.283
m8							0.023	0.015	0.128

A preliminary analysis shows that the effect of sex is not significant in the outcome equation of the models and it was removed. This further improve model identifiability in the context of the exclusion restriction criteria.

The intercept estimates of SSNM, SNM and TS models for the NDI scores at month 8 differ substantially, as expected from the simulation results (Table 4.4). Note that the treatment effect and measurements at month 4 are labeled 'physio' and 'prev' respectively in the table. Coefficient estimates in the outcome model vary less. As observed in the simulation study, the coefficients in the selection equations for SNM and TS are consistently larger than the SSNM model. In particular, the estimate of the skewness parameter ($\lambda = 1.537$) is statistically significant in the

Table 4.4: Fit of selection skew-normal model (SSNM), Selection-normal model (SNM), and Heckman two-step model to the NDI scores at 8 months.

	SSNM			SNM			TS		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
				Selection Equation					
int	0.208	0.177	0.239	0.835	0.100	0.000	0.818	0.115	0.000
age	0.021	0.005	0.000	0.024	0.006	0.000	0.025	0.006	0.000
sex(f)	0.309	0.126	0.014	0.335	0.129	0.009	0.383	0.152	0.012
				Outcome Equation					
int	-3.769	0.802	0.000	0.799	0.621	0.198	1.030	1.766	0.560
age	0.074	0.025	0.003	0.086	0.023	0.000	0.068	0.047	0.154
prev	0.678	0.035	0.000	0.687	0.035	0.000	0.708	0.035	0.000
physio	0.766	0.532	0.150	0.887	0.538	0.099	1.007	0.548	0.067
σ	7.723	0.563	0.000	6.166	0.292	0.000	5.703	2.036	0.005
ρ	0.758	0.174	0.000	0.802	0.072	0.000	0.474	0.641	0.460
λ	1.537	0.450	0.001	-	-	-	-	-	-

SSNM model. This implies that neglecting the influence of λ in the model, although it leads to the same qualitative conclusions for the covariate effects in the outcome equation (except age that is not significant at 5% level for the TS model), will lead to wrong predictive power of the model. The SSNM model has a better fit (log-likelihood = -1452.67) to the NDI data than the SNM model (log-likelihood = -1455.03). The SSNM is more general with the advantage of having good predictive power whether or not there is skewness in the data and, of course, has SNM as a special case.

A plot of fitted scores at month 8 against previous scores (month 4) for fixed values of age (40 years), sex and treatment are presented in Figures 4.5, 4.6 and 4.7 for the models in Table 4.4. A linear association (as expected) between measurements at months 8 and 4 is evident. The SSNM model provides a better fit to the data. To see this, consider a 40-year old male patient given physiotherapy with previous scores equals 11. His observed scores at month 8 is 12. However, the fitted values from SSNM, SNM and two-step models result in 12.61, 13.09 and 12.76 respectively.

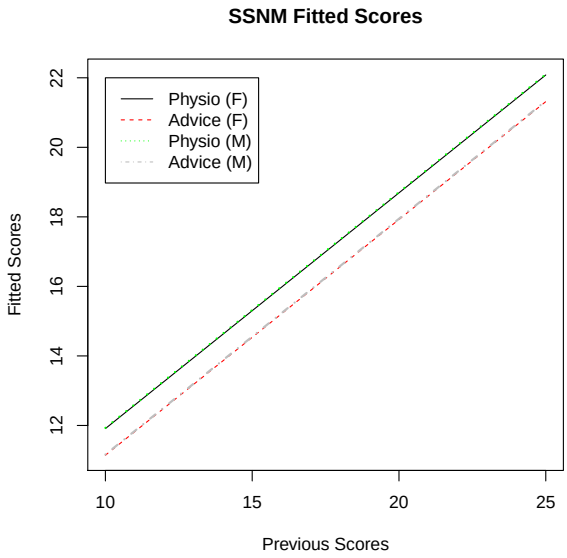

Figure 4.5: Fitted SSNM model.

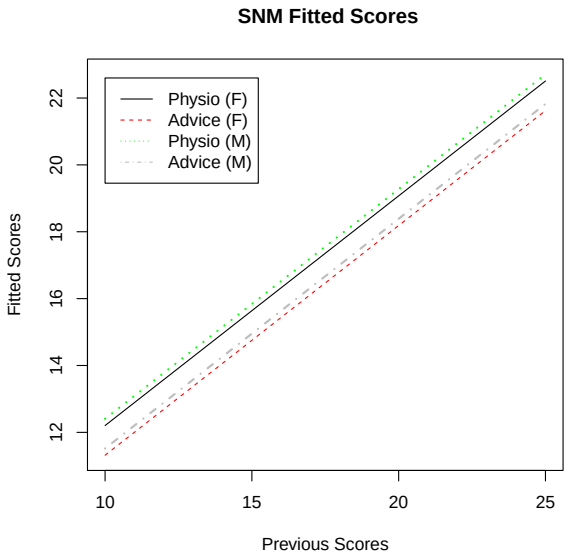

Figure 4.6: Fitted SNM model.

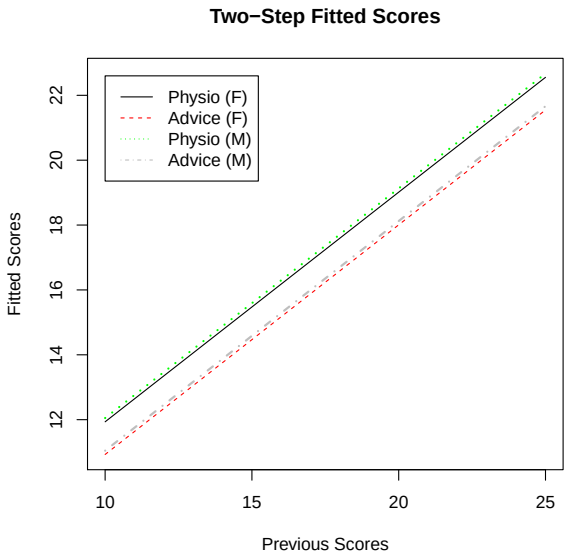

Figure 4.7: Fitted Two-step model.

4.2.4 Profile log-likelihood for the NDI scores

The introduction of extra parameters to a model, although leading to a more flexible model, comes at a cost of model identifiability in some cases. The profile log-

likelihood for the shape parameter of a univariate skew-normal distribution always has a stationary point at $\lambda = 0$. This problem is also visible in the SSNM model since it has the Azzalini's skew-normal distribution as its basis. To illustrate this, we examine the profile log-likelihood for the parameters λ, ρ and σ for the NDI scores. At $\lambda = 0$ in Figure 4.8, the profile log-likelihood has a stationary point, with log-likelihood value of -1455.03. The profile log-likelihood for ρ under the SSNM (see Figure 4.9) is flat in the neighborhood of zero, but less flat for SNM and may not affect inference about ρ. Although the Wald test in Table 4.4 shows that the correlation ρ is significant in the SSNM model (and also the SNM model), a likelihood ratio test for $\rho = 0$ gave a nonsignificant p-value (0.437). A similar test under the SNM model yielded a significant p-value (0.009). The discrepancy in the tests under SSNM model reflects the flat surface of the profile log-likelihood around zero. The profile log-likelihood for sigma in SSNM and SNM models (Figure 4.10) are much more regular, though again the SSNM profile is flatter.

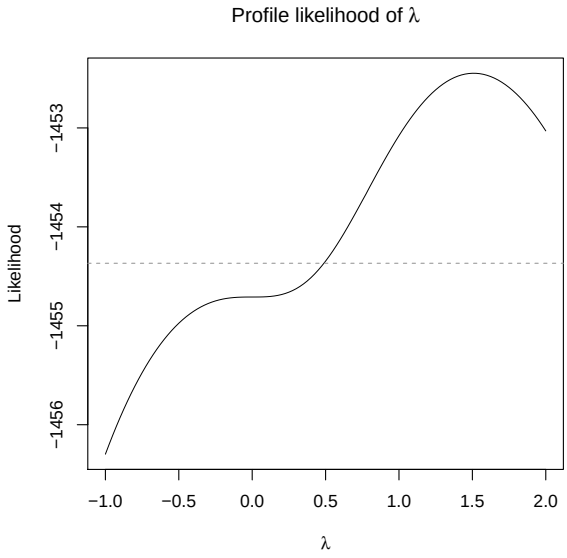

Figure 4.8: Profile log-likelihood for λ for the NDI scores (SSNM model).

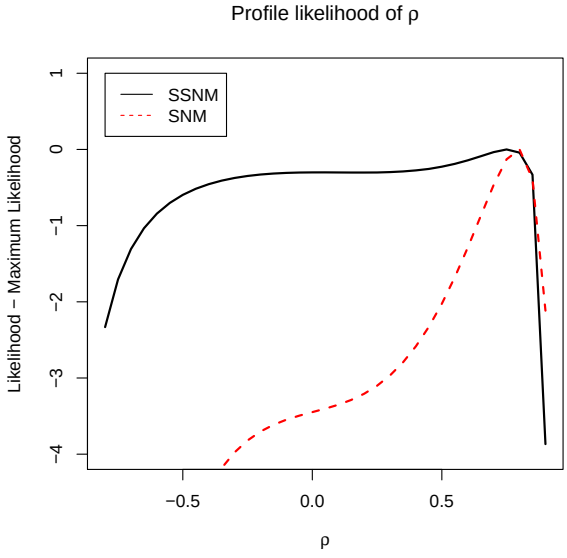

Figure 4.9: Profile log-likelihood for ρ for the NDI scores (SSNM & SNM models).

We assess the effects of profile log-likelihood surface flatness around zero on the parameter estimates when the SSNM model is fitted to the NDI scores at month 8 for fixed values of ρ i.e (-0.7, -0.5, 0, 0.5, 0.7, 0.8). There is a negative correlation between λ and ρ (Table 4.5).

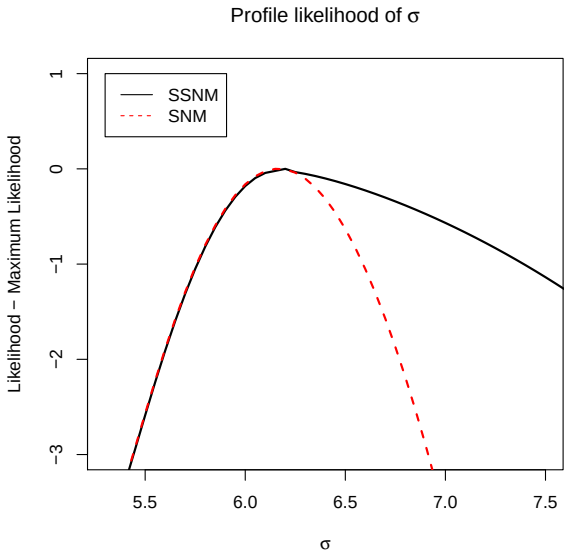

Figure 4.10: Profile log-likelihood for σ for the NDI scores (SSNM & SNM models).

The results under the SSNM model are consistent with $\rho = 0$, i.e. MAR, with the skewness in the response variable being intrinsic to the measured outcome, and not due to selection. Similarly, the TS model (with standard errors obtained from bootstrap) also supports the MAR assumption. However, the SNM model suggests the data are MNAR: if the outcome variable is normal in the population, informative missingness is required to explain the observed result.

A comparison of sample selection models in this chapter and the complete case analysis of chapter 3, using skew distributions, underscores the impact of the additional information due to binary regression in selection models. All the models fitted to the NDI scores at month 8 in chapter 3 showed that the treatment effect is significant (see Tables 3.3 and 3.4). However, the SSNM, SN and TS models, which correct for selection, show that the treatment effect is not significant.

Instead of modeling observed data in sample selection framework using skew-normal distribution, e.g. the ESN distribution, a model based on equation (3.2) should be preferred. This equation described the observed data satisfactorily because additional information about covariates can be incorporated in the model through $\gamma' x$. This model was fitted to the NDI scores at month 8 using restricted parameter space (i.e. $\rho = 0$ is excluded). The parameter estimates from the model (not shown here) gave results similar to the outcome model of the SNM model in Table 4.4, and the treatment effect is not significant. The main disadvantage of

Table 4.5: Fit of selection skew-normal model (SSNM) with 6 fixed value of ρ to the NDI scores at 8 months.

	$\rho = -0.7$ Estimate	$\rho = -0.5$ Estimate	$\rho = 0$ Estimate	$\rho = 0.5$ Estimate	$\rho = 0.7$ Estimate	$\rho = 0.8$ Estimate
			Selection Equation			
int	1.236	1.133	0.817	0.427	0.256	0.175
age	0.021	0.023	0.025	0.023	0.021	0.021
sex(f)	0.286	0.333	0.384	0.361	0.324	0.296
			Outcome Equation			
int	-2.855	-2.658	-2.806	-3.359	-3.674	-3.837
age	0.022	0.030	0.045	0.061	0.071	0.078
prev	0.672	0.675	0.677	0.679	0.679	0.676
physio	0.759	0.746	0.746	0.759	0.768	0.760
σ	8.846	8.139	7.559	7.554	7.664	7.786
λ	2.148	1.899	1.718	1.663	1.581	1.498
Loglik	-1453.98	-1453.27	-1452.97	-1452.90	-1452.71	-1452.72

modeling observed scores using (3.2) is that selection models cannot be consistently estimated.

Conditional normality and NDI scores at month 12

The SSNM, SNM and TS models are fitted to the NDI scores at the last measurement occasion, adjusting for previous measurements (m4 and m8). We also included age and sex in the model as biological factors that could predict non-response in the scores. Table 4.6 is the results of fitting the models to the NDI scores at month 12. The sample selection effect ($\rho \neq 0$) is not significant using the Wald test (p-value = 0.857) and the LRT affirm it with p-value = 0.863. The direct parameter from the Azzalini model (see Table 4.7) fitted to the data agrees closely with the parameters of the SSNM model in Table 4.6, further justifying that a complete case analysis may be sufficient to model the data. In addition, the skewness parameter is not significant both in the SSNM and the Azzalini skew-normal model. A LRT for $\lambda = 0$ in Table 4.6 also gave a non significant p-value (0.675). Although residual plots for the NDI scores (see Figure 2.5) showed that conditional normality is not tenable, adjusting for previous measurements at month 12 makes the residuals to be approximately normal.

Table 4.6: Fit of selection skew-normal model (SSNM), Selection-normal model (SNM), and Heckman two-step model to the NDI scores at 12 months.

	SSNM			SNM			TS		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
				Selection Equation					
int	0.240	0.306	0.433	0.293	0.105	0.005	0.293	0.105	0.005
age	0.020	0.005	0.000	0.020	0.005	0.000	0.020	0.005	0.000
sex(f)	0.387	0.134	0.004	0.387	0.134	0.004	0.387	0.135	0.005
				Outcome Equation					
int	-2.391	2.016	0.236	-1.215	1.583	0.443	-1.245	2.633	0.637
age	0.017	0.031	0.572	0.018	0.032	0.572	0.019	0.046	0.688
physio	-0.924	0.540	0.088	-0.901	0.539	0.095	-0.900	0.551	0.103
base	0.092	0.046	0.046	0.096	0.046	0.036	0.096	0.047	0.042
m4	0.215	0.050	0.000	0.217	0.049	0.000	0.217	0.055	0.000
m8	0.618	0.051	0.000	0.620	0.050	0.000	0.620	0.054	0.000
σ	5.112	0.770	0.000	4.570	0.262	0.000	4.574	2.300	0.047
ρ	0.114	0.636	0.857	0.138	0.600	0.819	0.150	0.591	0.800
λ	0.694	0.583	0.235	-	-	-	-	-	-
Loglik	-1132.24			-1132.33					

Table 4.7: Complete cases with Azzalini Skew-normal errors and Normal errors.

	Direct Param			OLS Param		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value
int	-2.192	1.620	0.177	-0.899	0.777	0.248
age	0.013	0.020	0.515	0.013	0.021	0.544
physio	-0.929	0.540	0.087	-0.901	0.544	0.099
base	0.092	0.046	0.0460	0.096	0.046	0.039
m4	0.214	0.050	0.000	0.216	0.050	0.000
m8	0.618	0.051	0.000	0.620	0.051	0.000
σ	5.128	0.773	0.000	4.595	0.112	0.000
λ	0.710	0.564	0.209			

4.3 Possible extensions of the SSNM models

We present a brief overview of two extensions of the SSNM model that can be of practical interest. A multivariate extension is given in order to emphasis the use of the model in a longitudinal framework, while extension to model skewness and heavy-tail simultaneously is given to emphasise the generality of sample selection models.

4.3.1 Multivariate extension of the SSNM model

The model proposed in this chapter, although it made explicit assumption about non-response mechanism, is deficient in that it cannot capture average evolution of treatment effect which is the ultimate goal of any longitudinal study. A multivariate extension of this model can readily be developed using the CSN distribution. We maintain the general formulation (2.17).

Suppose a $p \times 1$ random vector $\mathbf{Y}$ of outcomes follows a SN distribution with a $p \times 1$ location vector $\boldsymbol{\beta}'x$, $p \times p$ symmetric positive definite dispersion matrix Ω and $p \times 1$ vector of skewness parameter $\boldsymbol{\lambda}$. That is, $\mathbf{Y} \sim SN_p(\boldsymbol{\beta}x, \boldsymbol{\Omega}, \boldsymbol{\lambda})$. Suppose further that S is a selection mechanism which is normally distributed with mean γx and variance 1. This implies that the selection patterns across the p-dimensional outcomes are the same. Using the approach of (4.6), the joint distribution of the outcomes and the selection process can be written as

$$\begin{pmatrix} \mathbf{Y} \\ S \end{pmatrix} \sim CSN_{p+1,1}\Big\{\boldsymbol{\mu} = (\boldsymbol{\beta}'x, \gamma'x), \Sigma = \begin{pmatrix} \Omega & \rho\Omega^{1/2} \\ \rho\Omega^{1/2} & 1 \end{pmatrix}, D = (\boldsymbol{\lambda}'\Omega^{1/2}, 0), \boldsymbol{\nu} = 0, \Delta = 1\Big\},$$

where ρ is the correlation between each element of $\mathbf{Y}$ and S. When $\rho = 0$, there is no selection, as in the case of the SSNM model. The distribution of $S|\mathbf{Y}$ in CSN form is

$$CSN_{1,1}\Big\{\gamma'x + \rho\Omega^{-1/2}(\mathbf{y} - \boldsymbol{\beta}'x), 1 - \rho^2, 0, -\boldsymbol{\lambda}'\Omega^{-1/2}(\mathbf{y} - \boldsymbol{\beta}'x), 1\Big\}. \tag{4.18}$$

Notice that all the matrices are conformable for multiplication. Equation (4.18) is an ESN distribution. Since the skewness parameter is zero, it turns out that the distribution is a normal distribution. So

$$P(S^\star > 0|\mathbf{Y}, x) = \Phi\bigg(\frac{\gamma'x + \rho\Omega^{-1/2}(\mathbf{y} - \boldsymbol{\beta}'x)}{\sqrt{1-\rho^2}}\bigg).$$

The normalizing function $P(S^\star > 0)$ has a CSN representation

$$CSN_{1,1}\Big\{\gamma'x, 1, \boldsymbol{\lambda}'\rho, 0, 1 + \boldsymbol{\lambda}'\boldsymbol{\lambda}(1-\rho^2)\Big\}.$$

This implies

$$P(S^\star > 0) = \Phi_{SN}\bigg(\gamma'x; 0, 1, \frac{-\boldsymbol{\lambda}'\rho}{\sqrt{1 + \boldsymbol{\lambda}'\boldsymbol{\lambda}(1-\rho^2)}}\bigg),$$

which is a univariate skew-normal distribution with skewness parameter $-\boldsymbol{\lambda}'\rho/\sqrt{1 + \boldsymbol{\lambda}'\boldsymbol{\lambda}(1-\rho^2)}$.

The continuous component of the multivariate SSNM model is therefore

$$\frac{2\phi_p\Big(\mathbf{y};\boldsymbol{\beta}'x,\Omega\Big)\Phi\Big(\boldsymbol{\lambda}'\Omega^{-1/2}(\mathbf{y}-\boldsymbol{\beta}'x)\Big)\Phi\left(\frac{\gamma'x+\rho\Omega^{-1/2}(\mathbf{y}-\boldsymbol{\beta}'x)}{\sqrt{1-\rho^2}}\right)}{\Phi_{SN}\left(\gamma'x;0,1,\frac{-\boldsymbol{\lambda}'\rho}{\sqrt{1+\boldsymbol{\lambda}'\boldsymbol{\lambda}(1-\rho^2)}}\right)}. \tag{4.19}$$

If $\rho = 0$ in (4.19), the multivariate skew-normal distribution is recovered. The SSNM model can be derived from this generalization when the dimension p of the outcome equation is 1. Similarly, the hidden truncation formulation of this model is straightforward if we make use of equation 5.5 of Arnold and Beaver (2002).

The complete sample selection density of the multivariate SSNM has (4.19) as its continuous component. The selection part is a binary regression with the skew-normal link. One major challenge in using this model is how to model the covariance structure over time and the estimation of all the skewness parameters from the available data. We will examine the impact of boundedness of the scores on the covariance structure in our future work. In the same vein, modeling simultaneously the two prominent deviations from normality assumption (skewness and heavy-tail) may be of interest. Although this is beyond the scope of this thesis, we show in the next section that a model with underlying bivariate skew-t distribution can be derived using the same approach that we used for the development of the SSNM model.

4.3.2 Sample selection model with skew-t distribution

There is a noticeable pattern in the construction of the models in this chapter. When the underlying distributional assumption is bivariate normal, the continuous component of the sample selection density is from the ESN distribution. Marchenko and Genton (2012) used a bivariate-t distribution and the continuous component is an extended skew-t (EST) distribution (Arellano-Valle and Genton (2010)). When the underlying distribution is no longer elliptical, as we've shown here, the continuous component of sample selection density is still in the form given by (3.1) but the derivation is more complicated. The model that we derive in this chapter used the flexibility of the CSN distribution to construct the continuous component of sample selection density.

A more general sample selection model can be described using an underlying bivariate skew-t distribution. The advantage of this model is that it has the Heckman (1976), Marchenko and Genton (2012), and the SSNM models as special cases. We expect the model to capture skewness, heavier tails than the normal, mixtures of

normal distributions, and some contaminated normal data sets. We define next a skew-t distribution.

Definition 8. *A p-dimensional random vector* $\mathbf{Y}$ *is said to have a skew-t distribution if its PDF is of the form*

$$f(\mathbf{y}) = 2t_p(\mathbf{y};\eta)T_1\left(\boldsymbol{\lambda}'\omega^{-1}(\mathbf{y}-\boldsymbol{\xi})\left[\frac{\eta+p}{Q+\eta}\right]^{1/2};\eta+p\right), \tag{4.20}$$

where t_p *is the density of a p-dimensional t variate with* η *degrees of freedom:*

$$t_p(\mathbf{y};\eta) = \frac{\Gamma((\eta+p)/2)}{|\Omega|^{1/2}(\pi\eta)^{p/2}\Gamma(\eta/2)}(1+Q/\eta)^{(\eta+p)/2}, \tag{4.21}$$

where

$$Q = (\mathbf{y}-\boldsymbol{\xi})'\Omega^{-1}(\mathbf{y}-\boldsymbol{\xi}).$$

The scalar parameter $\eta > 0$ denotes the degrees of freedom of the multivariate t-distribution and Γ is the gamma function. The p-dimensional vector $\boldsymbol{\xi}$ is a location parameter, Ω is a $p\times p$ covariance matrix and $\omega = \text{diag}(\Omega)^{1/2}$. The skewing function, $T_1(.;\eta+p)$, is a univariate t distribution function with $\eta+p$ degrees of freedom. The p-dimensional vector $\boldsymbol{\lambda}$ controls the skewness. A tool for the construction of sample selection model with underlying bivariate skew-t distribution can easily be developed.

Consider equation (2.18) but with an underlying multivariate t-distribution. That is

$$\begin{aligned}\mathbf{Y}^\star &= \boldsymbol{\mu}+\boldsymbol{\varepsilon}_1, \quad \boldsymbol{\varepsilon}_1 \sim t_p(\mathbf{0};\Omega,\eta)\\ \mathbf{S}^\star &= -\boldsymbol{\nu}+D\boldsymbol{\mu}+\boldsymbol{\varepsilon}_2,\ , \quad \boldsymbol{\varepsilon}_2 \sim t_q(\mathbf{0};\Delta,\eta),\end{aligned} \tag{4.22}$$

where $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are independent random vectors, and $D(q\times p)$ is an arbitrary matrix, $\boldsymbol{\mu}\in\mathbb{R}^p$, $\boldsymbol{\nu}\in\mathbb{R}^q$, $\Delta(q\times q)>0$, and $\eta>0$. The joint distribution of $\mathbf{Y}^\star$ and $\mathbf{S}^\star$ is

$$\begin{pmatrix}\mathbf{Y}^\star\\ \mathbf{S}^\star\end{pmatrix} \sim t_{p+q}\left(\begin{pmatrix}\boldsymbol{\mu}\\ -\boldsymbol{\nu}\end{pmatrix},\begin{pmatrix}\Omega & \Omega D'\\ D\Omega & \Delta+D\Omega D'\end{pmatrix},\eta\right).$$

The conditional density $(\mathbf{y}^\star|\mathbf{s}^\star>\mathbf{0})$ after some algebra yields a closed skew-t (CST) distribution $CST_{p,q}(\boldsymbol{\mu},\Sigma,D,\boldsymbol{\nu},\Delta,\eta)$ with density

$$f(\mathbf{y}) = \frac{1}{T_q(\mathbf{0};\boldsymbol{\nu},\Delta+D\Omega D',\eta)}t_p(\mathbf{y};\boldsymbol{\mu},\Omega,\eta)T_q(D(\mathbf{y}-\boldsymbol{\mu});\boldsymbol{\nu},\Delta,\eta+p),$$

where $T_q(.;\boldsymbol{\mu},\Psi,\eta)$ is the CDF of a q-dimensional t-distribution with mean $\boldsymbol{\mu}\in\mathbb{R}^q$, $q\times q$ covariance matrix Ψ and η degrees of freedom. The CST distribution can be reparametrized to form the so-called unified skew-t (SUT) distribution introduced in Arellano-Valle and Genton (2010). As it turns out, when $\eta\to\infty$ in the CST distribution, CSN distribution is obtained.

If we assume a bivariate skew-t distribution for the outcome and selection equation, but restrict the skewness parameter to zero in the selection equation, we can develop a new class of sample selection model. The continuous component of the sample selection model is a form of the univariate extended skew-t distribution, EST (see Arellano-Valle and Genton (2010)), but with an additional skewness parameter. Arellano-Valle and Genton (2010) gave analytic proof that the EST distribution, unlike the ESN, does not have a stationary point at $\lambda=0$. We expect the additional skewness parameter in the EST distribution not to induce stationarity at $\lambda=0$ in the model, and thus produce a more stable parameter estimates than the SSNM model. In addition, the selection equation is a binary regression with the skew-t link (see Kim (2002)).

4.4 Summary

We introduced a sample selection model with underlying bivariate skew-normal distribution which we called selection skew-normal model (SSNM). This model is more flexible than the conventional sample selection model since it has an extra parameter that regulates skewness and has conventional sample selection model as a special case. Its moment estimator was derived using the link between skew models arising from selection and hidden truncation formulation of skew models. The moment estimator was shown to extend Heckman two-step method. Maximum likelihood estimation was considered using a Monte Carlo study to compare the model with conventional sample selection models with moderate correlation ($\rho=0.5$) and varying degree of skewness between 0 and 5. We also fixed λ to be 1 and 2, and considered the effect of varying the correlation ρ under the exclusion restriction criteria (see Tables A.1 and A.2 in Appendix A). The simulation showed that the SSNM model outperforms the conventional sample selection models for all the skewness parameters considered. The conventional sample selection model has a negligible advantage when $\lambda=0$ with smaller bias in the intercept of outcome equation. We also noted that the conventional sample selection model breaks down as λ increases to 5 (which is almost a folded normal distribution) and the SSNM works well if it converges. The model is very promising even in the absence of exclusion restriction criteria.

In addition, the model has good estimates of the intercept both in the selection and outcome equations and hence will give better predictions even when the underlying process is bivariate normal. This model should perform better than the conventional sample selection model in modeling heavier tailed data.

The model presented here is very simple to use and the likelihood function can be easily coded in R software. Starting values can be obtained using the two-step method (TS). However, we recommend obtaining a starting value for λ by fitting the Azzalini skew-normal model (direct parametrization) to complete cases with the intended covariates for the outcome equation. Further, the optimization routine used was BFGS in R software but other numerical maximization algorithms can be used as well (although we do not recommend the use of Nelder-Mead optimization method which appears not to work well with the CDF of Azzalini's skew-normal distribution). We recommend the use of profile likelihood for λ to avoid convergence to local maxima.

On the issue of model identification, the model is well identified in the sense that for any $\Theta_1 \neq \Theta_2$, $f(y, \Theta_1) \neq f(y, \Theta_2)$, where Θ_1 and Θ_2 are model parameters. Further, the observed information matrix is non-singular (see section A.1 in the Appendix A for the elements of the observed information matrix). This is usually the case with sample selection models since additional information comes into the model through the selection process. However, in the absence of exclusion restriction and with λ approaching infinity, the model is weakly identified. It is noteworthy that inference about λ and ρ may not be feasible when the two parameters equal zero. This is not related to the identification of the model parameters but the stationarity of profile log-likelihood of λ and ρ at zero. In addition, the observed information matrix is not singular when either λ or ρ is zero. Since the stationarity problem of λ was inherited from the underlying Azzalini's skew-normal distribution used, a more flexible skew distribution (not based on the perturbation of normal kernel) can be used and the use of sinh-arcsinh distribution of Jones and Pewsey (2009) will be considered in chapter 6.

We noted that model (4.10) is more general than the one presented here. However, it is computationally complicated. Apart from this, the parameter ρ is no longer adequate to capture the underlying association. The model therefore needs to be re-parameterized using correlation curves. In addition, since the marginal distribution of the observed data are known to be skew, copula based sample selection models can be used. A bivariate Gaussian copula, similar to Lee (1983) model, but with skew-normal and normal margins was compared with the SSNM model and the results were shown to be similar. The stationarity of profile likelihood for λ

at $\lambda = 0$ persisted and the surface of the profile likelihood of ρ remained flat in the neighborhood of zero. These give further credence to the fact that the stationarity problem is not peculiar to the SSNM model but to the underlying Azzalini skew-normal distribution used.

To apply this model in practice, we recommend that the model is fitted in conjunction with the conventional sample selection model. This can be used to assess the degree of departure from symmetry. The model could be of benefit in clinical trials and it has prospects in fields where observational studies are conducted (econometrics, psychology, politics) and respondents need to complete questionnaires.

Chapter 5

A Unified Approach to Multilevel Sample Selection Models

The models proposed for the analysis of the NDI scores so far have not distinguished between the two levels of non-response present in the data set. Unit non-response occur when a subject declines participation in a study, and item non-response occur when questions are skipped. We can regard the observed outcomes as the result of a two level selection process. That is, both unit and item non-response simultaneously affect the outcome of interest and both types of non-response are potentially correlated. This distinction can be used to study factors that affect the two non-response types independently and jointly. In this chapter, we consider the observed outcomes as realizations from a non-truncated marginal of a truncated multivariate normal distribution. The resulting density for the outcome is the continuous component of the sample selection density, and has links with the CSN distribution. The CSN distribution provides a framework which simplifies the derivation of the conditional expectation and variance of the observed data. We use this to generalize the moment based Heckman's two-step method to a multilevel sample selection model. A simulation study is used to study finite sample performances of the moment and likehood based estimators of the model. In addition, since the NDI scores are skew, we propose an extension of the SSNM model of chapter 4, with skew outcomes and two normally distributed selection processes.

5.1 Multilevel Sample Selection Models

Multilevel sample selection arises when more than one selection process affects the outcome of interest in a study. These models have been discussed in the literature in various forms. Poirier (1980) investigated random utility models in which observed binary outcomes do not reflect the binary choice of a single decision-maker, but rather the joint unobserved binary choices of two decision-makers. This model was further developed by Ham (1982). A slight modification of this model was considered in Luca and Peracchi (2006) in which an extension of Poirier (1980) model was used to jointly analyze items and unit non-response in survey data. Further application of multilevel selection models in cross-sectional settings can be found in Bellio and Gori (2003), Arendt and Holm (2006) and Rosenman et al. (2010).

Recall the regression model given in section 3.1, that is

$$Y_i^\star = \beta' x_i + \sigma\varepsilon_{1i}, \quad i = 1, \ldots, N,$$

as regression model of interest, but now with n possible selection processes (not necessarily hierarchical) given as

$$\begin{cases} S_{1i}^\star & = \alpha_1' x_i + \varepsilon_{2i} \\ S_{2i}^\star & = \alpha_2' x_i + \varepsilon_{3i} \\ \vdots & \\ S_{ni}^\star & = \alpha_n' x_i + \varepsilon_{(n+1)i}, \end{cases}$$

where $S_{1i} = I(S_{1i}^\star > 0)$, $S_{2i} = I(S_{2i}^\star > 0), \ldots, S_{ni} = I(S_{ni}^\star > 0)$. The usable observations are the $Y_i = Y_i^\star * S_{1i} * S_{2i} \cdots * S_{ni}$ with density $f(y_i|x_i, S_{1i} = 1, S_{2i} = 1, \ldots, S_{ni} = 1)$. This density is the continuous component of the multilevel sample selection density. The discrete component is determined by the marginal distribution of the selection mechanisms. Unlike in single selection process, the binary regression is determined by the nature of the selection process.

When multilevel selection models are mentioned in the literature (econometric literature in particular), what usually comes to mind is a two-level selection process. This has an outcome equation (binary or continuous outcomes) and two selection equations with trivariate Gaussian error distribution assumption. At the end, a two-level extension of the Heckman two-step method is derived and used to analyze the observed data. However, there are cases where more than two selection processes can affect the outcome of interest. In some of these cases, the selection mechanisms are combined to make the model more manageable and the complicated

algebra required to write more than two-level Heckman selection method is avoided in the process. In principle, the observed outcomes follow the CSN distribution which can easily be constructed using the link between sample selection and skew distributions of section 2.2 (hidden truncation or conditioning). Properties of the CSN distribution can then be used to generalize multilevel sample selection models to any number of selection processes in a straightforward way.

Without loss of generality, we use a two-level selection model to illustrate the unification of multilevel sample selection problems into a distributional framework. We begin by quantifying the overall effect of two nested non-response mechanisms on the regression outcome of interest.

5.2 Mathematical formulation of the Model

In section 3.1, we showed that the continuous component of sample selection density is an ESN density. Since the CSN density is a unifying class for the Azzalini (1985) SN family, the ESN is necessarily a member. In fact, we wrote an ESN density in a CSN form in section 3.4.1. Thus equation (3.2), the continuous component of Heckman (1976) sample selection density, can be written as

$$f(y|x, S=1) = \frac{\phi\Big(y; \beta'x, \sigma^2\Big)\,\Phi\Big(\frac{\rho}{\sigma}\Big(y-\beta'x\Big); -\gamma'x, 1-\rho^2\Big)}{\Phi\Big(0; -\gamma'x, 1\Big)},$$

which has CSN form

$$(Y|x, S=1) \sim CSN_{1,1}\Big(\beta'x, \sigma^2, \frac{\rho}{\sigma}, -\gamma'x, 1-\rho^2\Big).$$

One can as well make an educated guess that the continuous component of multilevel sample selection density is a CSN density using equation (2.18). What we need to determine is the structure of the density in a sample selection settings. We first look at statistical bias in two-level hierarchical selection problem and show when the non-response processes is ignorable

5.2.1 Statistical bias in two-level sample selection problem

In this section, we present an expression that quantifies the overall non-response bias in two-level sample selection model. The non-response mechanism under which the bias vanishes is also described in a manner similar to the one discussed in Luca and Peracchi (2006). The model is developed by assuming the two-level selection equa-

tions correspond to unit and item non-response. We begin by extending equations (4.1) and (4.2) with an additional selection equation,

$$S_{2i}^{\star} = \alpha' x_i + \varepsilon_{3i}, \quad i = 1, \ldots, N. \tag{5.1}$$

Since we have two selection equations, we can take $S_{1i} = I(S_{1i}^{\star} > 0)$ and $S_{2i} = I(S_{2i}^{\star} > 0)$. The usable observations are on $Y_i = Y_i^{\star} S_{1i} S_{2i}$.

Now, the interest is to estimate the conditional mean function of a random outcome using data from a clinical study. Suppose at each time points, N patients are expected to respond, then unit non-response may reduce the number of patients to $N_1 < N$ responding units. Further, non-response at item level may reduce effective number of observations to $N_2 < N_1$. The loss of information due to missing observations results in efficiency loss relative to the ideal situation of complete response.

It is logical to consider at each time point a sequential framework where patients are first observed before they decide to answer specific item of the questionnaire. Let the indicator of unit response be S_1, which is always observed, while the indicator of item response be S_2 which is observed conditional on S_1 being present. Since the observations are present when the indicators are greater than zero, we can then describe the response process by $\pi_0 = \Pr\{S_1 = 0\}$ and $\pi_{0|1} = \Pr\{S_2 = 0 | S_1 = 1\}$ representing the probability of unit non-response and the probability of item non-response conditional on unit response respectively. Since Y is the outcome of interest, we have using, the law of iterated expectations,

$$E(Y|S_1 = 1) - E(Y) = \pi_0[E(Y|S_1 = 1) - E(Y|S_1 = 0)]. \tag{5.2}$$

In addition,

$$E(Y|S_1 = 1) = E(Y|S_1 = 1, S_2 = 1) + \pi_{0|1}[E(Y|S_1 = 1, S_2 = 0) - E(Y|S_1 = 1, S_2 = 1)]. \tag{5.3}$$

The difference between the conditional mean of Y for the fully responding patients and the unconditional mean of Y for the complete response is the overall non-response bias and is given as $E(Y|S_1 = 1, S_2 = 1)$ - $E(Y)$. Substituting (5.3) into (5.2) and rearranging gives

$$\begin{aligned} E(Y|S_1 = 1, S_2 = 1) - E(Y) = \pi_0[E(Y|S_1 = 1) - E(Y|S_1 = 0)] + \\ \pi_{0|1}[E(Y|S_1 = 1, S_2 = 1) - E(Y|S_1 = 1, S_2 = 0)]. \end{aligned}$$

Generally, the overall bias has two separate components that are proportional to the probabilities of unit and item non-response respectively. There are 3 ways by which the bias can be zero in the above equation. If there is neither unit nor item non-response ($\pi_0 = \pi_{0|1} = 0$), if both unit and item non-response are MAR ($E(Y|S_1 = 1) = E(Y|S_1 = 0)$ and $E(Y|S_1 = 1, S_2 = 1) = E(Y|S_1 = 1, S_2 = 0)$), and when the bias terms due to unit and item non-response have opposite sign and offset each other. Next, we consider a model that removes this bias.

5.2.2 Two-level selection models

Recall that the hidden truncation method of Arnold and Beaver (2002) and skew distributions arising from selection of Arellano-Valle et al. (2006) were used to derive the continuous component of sample selection density in section 3.1. The same approach can be used here although the derivation of the conditional mean and variance is complicated when there is more than one selection equation. The moment generating function (mgf) of the CSN distribution can be used to simplify this.

Hidden truncation method

Suppose $f(y, s_1, s_2)$ is the density of a trivariate normal random variable with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)'$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & \sigma\rho_{12} & \sigma\rho_{13} \\ \sigma\rho_{12} & 1 & \rho_{23} \\ \sigma\rho_{13} & \rho_{23} & 1 \end{pmatrix}. \tag{5.4}$$

Suppose further that $W = (Y, S_1, S_2)'$ has joint density

$$\begin{cases} f(\mathbf{w}) & = \frac{1}{C}\frac{1}{\sqrt{(2\pi)^3|\Sigma|}} e^{-1/2(\mathbf{w}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})}, \quad \mathbf{w} \in R \\ & = 0, \text{ otherwise} \end{cases}$$

where R is a rectangle in 3-space; R: $-\infty < y < \infty$, $c_{s_1} < s_1 < \infty$ and $c_{s_2} < s_2 < \infty$. C is a normalizing constant (necessary to ensure that the density function integrates to 1) given by

$$C = \int_R \frac{1}{C}\frac{1}{\sqrt{(2\pi)^3|\Sigma|}} e^{-1/2(\mathbf{w}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})} \mathrm{d}\mathbf{w}.$$

This implies (Y, S_1, S_2) has a truncated trivariate normal distribution. S_1 and S_2 are truncated below at c_{s_1} and c_{s_2} respectively. We are interested in the marginal distribution of Y, which is the only non-truncated random variable in this formulation.

Using Cartinhour (1990), we can write the required density as,

$$f(y) = \frac{1}{C} e^{-1/2(\frac{y-\mu_1}{\sigma^2})^2} \int_{c_{s_1}}^{\infty} \int_{c_{s_2}}^{\infty} \frac{1}{\sqrt{(2\pi)^2 |A_{\neg y}^{-1}|}} e^{-1/2(\mathbf{w}_{\neg y} - \mathbf{m}(y))' A_{\neg y} (\mathbf{w}_{\neg y} - \mathbf{m}(y))} \mathrm{d}\mathbf{w}_{\neg y}, \tag{5.5}$$

where $\mathbf{w}_{\neg y} = (s_1, s_2)'$, $A_{\neg y}^{-1} = \Sigma_2^\star = \begin{pmatrix} 1-\rho_{12}^2 & \rho_{23}-\rho_{12}\rho_{13} \\ \rho_{23}-\rho_{12}\rho_{13} & 1-\rho_{13}^2 \end{pmatrix}$ (this is the inverse of the submatrix of the inverse of Σ when the row and column corresponding to y is deleted), and $\mathbf{m}(y)$ is defined as $\mathbf{m}(y) = \mu_{\neg 1} + (y - \mu_1/\sigma^2)\mathbf{k}$; with $\mu_{\neg 1} = (\mu_2, \mu_3)$, and $\mathbf{k} = (\sigma\rho_{12}, \sigma\rho_{13})'$. We determine C and the double integral in equation (5.5).

Now, C can be written as a noncentral normal integral

$$\Phi_3 \left(\begin{pmatrix} -\infty \\ c_{s_1} \\ c_{s_2} \end{pmatrix}, \begin{pmatrix} \infty \\ \infty \\ \infty \end{pmatrix}, \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}; \Sigma \right).$$

When the above is centralized, we have

$$\Phi_3 \left(\begin{pmatrix} -\infty \\ c_{s_1} - \mu_2 \\ c_{s_2} - \mu_3 \end{pmatrix}, \begin{pmatrix} \infty \\ \infty \\ \infty \end{pmatrix}; \Sigma \right) = \Phi_2 \left(\begin{pmatrix} 0 \\ 0, \end{pmatrix}, \begin{pmatrix} c_{s_1} - \mu_2 \\ c_{s_2} - \mu_3 \end{pmatrix}; \Sigma_2 \right), \tag{5.6}$$

where $\Sigma_2 = \begin{pmatrix} 1 & \rho_{23} \\ \rho_{23} & 1 \end{pmatrix}$. Using properties of multivariate normal cumulative distribution function and the definition of $\mathbf{m}(y)$, the double integral reduces to

$$\Phi_2 \left(\begin{pmatrix} \sigma\rho_{12} \\ \sigma\rho_{13} \end{pmatrix} \left(\frac{y - \mu_1}{\sigma^2} \right); \begin{pmatrix} c_{s_1} - \mu_2 \\ c_{s_2} - \mu_3 \end{pmatrix}, \Sigma_2^\star \right). \tag{5.7}$$

The required density is derived when equations (5.6) and (5.7) are substituted in equation (5.5). The PDF is

$$\frac{\phi(y; \mu_1, \sigma^2) \Phi_2(D(y - \mu_1); \boldsymbol{\nu}, \Sigma_2^\star)}{\Phi_2(\mathbf{0}; \boldsymbol{\nu}, \Sigma_2)}, \tag{5.8}$$

where $\mathbf{0} = (0, 0)'$, $D = (\rho_{12}/\sigma, \rho_{13}/\sigma)'$, and $\boldsymbol{\nu} = (c_{s_1} - \mu_2, c_{s_2} - \mu_3)'$. It is easy to see that $\Sigma_2 = \Sigma_2^\star + D\sigma^2 D'$, and thus (5.8) belongs to the closed skew-normal (CSN) family.

A plot of the PDF given by (5.8) is shown in Figure 5.1. The 'CSN(Normal)' represents the normal distribution as a special case of the CSN distribution. The

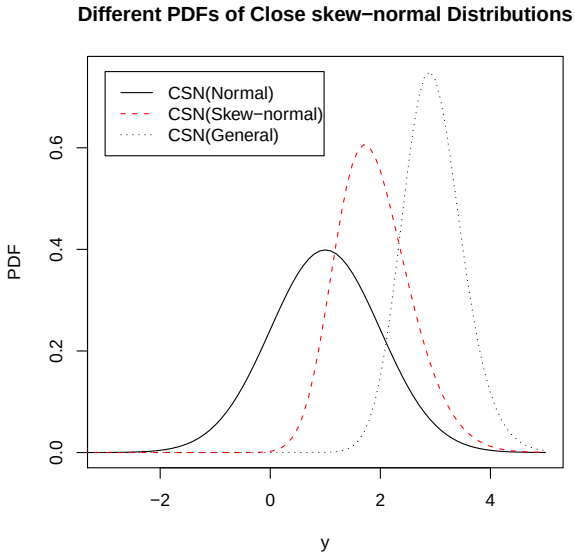

Figure 5.1: Comparison of Close skew-normal densities

parameters are $\mu_1 = 1$, $\sigma = 1$, $D = (0,0)'$, $\boldsymbol{\nu} = (0,0)'$, and $\Sigma_2^\star = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. The 'CSN(Skew-normal)' is a skew-normal equivalence of CSN distribution with $D = (1,2)'$, and other parameters kept as in the normal case. The more general form of the CSN is marked as 'CSN(General)' with $\boldsymbol{\nu} = (-2,4)'$ and other parameters kept as in the skew-normal, and it appears symmetric in Figure 5.1. The more general CSN can be more or less skew depending on its parameters. Thus, the need for model formulation in the general CSN family.

Arellano-Valle et al. (2006) equivalence of (5.8) can be obtained by restricting c_{s_1} & c_{s_2} to be zero, and using regression parametrization $\mu_1 = \beta' x$, $\mu_2 = \gamma' x$ and $\mu_3 = \alpha' x$. We then obtain,

$$\frac{\phi(y;\beta' x,\sigma^2)\Phi_2\left(D(y-\beta' x);\begin{pmatrix}-\gamma' x\\ -\alpha' x\end{pmatrix},\Sigma_2^\star\right)}{\Phi_2\left(\begin{pmatrix}0\\0\end{pmatrix};\begin{pmatrix}-\gamma' x\\ -\alpha' x\end{pmatrix},\Sigma_2\right)}. \tag{5.9}$$

The mathematical rigor in the derivation of (5.9) can be avoided using skew distributions arising from selection.

Skew distributions arising from selection method

Suppose we consider (4.1), the outcome equation and (4.2) and (5.1), the selection equations such that the error terms are distributed normally with means zero and covariance matrix given by (5.4). Then,

$$\begin{pmatrix} Y \\ S_1 \\ S_2 \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \beta' x \\ \gamma' x \\ \alpha' x \end{pmatrix} \begin{pmatrix} \sigma^2 & \sigma\rho_{12} & \sigma\rho_{13} \\ \sigma\rho_{12} & 1 & \rho_{23} \\ \sigma\rho_{13} & \rho_{23} & 1 \end{pmatrix} \right).$$

Now, (3.1) can be generalized to a two-level selection model as

$$f(y|x, S_1 = 1, S_2 = 1)) = \frac{f(y|x)P(S_1 = 1, S_2 = 1|y, x)}{P(S_1 = 1, S_2 = 1)}. \tag{5.10}$$

The quantity $f(y|x)$ is a proper PDF with a skewing function $P(S_1 = 1, S_2 = 1|y, x)$ and a normalizing function $P(S_1 = 1, S_2 = 1)$ to ensure that the LHS (left-hand side) of (5.10) integrates to 1. The marginal distribution of Y is $f(y|x) = \phi(y; \beta' x, \sigma^2)$. Similarly,

$$P(S_1 = 1, S_2 = 1) = 1 - \Phi_2(-\gamma' x, -\alpha' x; \rho_{23}) = \Phi_2(\gamma' x, \alpha' x; \rho_{23}).$$

Using the conditional distribution properties of the normal distribution, $P(S_1 = 1, S_2 = 1|y, x)$ becomes

$$\Phi_2\Big(D(y - \beta' x); \begin{pmatrix} -\gamma' x \\ -\alpha' x \end{pmatrix}, \Sigma_2^\star \Big),$$

where D and $\Sigma_2^\star$ are as defined in section 5.2.2. When appropriate substitutions are made in equation (5.10), the resulting density becomes:

$$\frac{\phi(y; \beta' x, \sigma^2)\Phi_2\left(\frac{\gamma' x + \rho_{12}(\frac{y-\beta' x}{\sigma})}{\sqrt{1-\rho_{12}^2}}, \frac{\alpha' x + \rho_{13}(\frac{y-\beta' x}{\sigma})}{\sqrt{1-\rho_{13}^2}}; \frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}} \right)}{\Phi_2(\gamma' x, \alpha' x; \rho_{23})}, \tag{5.11}$$

which is the standardized version of (5.9). Equation (5.11) is equivalent to equation 10 given in Ahn (1992).

In general, a CSN density is the continuous component of the multilevel sample selection density. In the bivariate case, it is given by equation (5.11). The discrete component of the log-likelihood function can be described by a bivariate probit model since the marginal distribution of the selection equation is a bivariate

normal distribution. Roughly speaking, the normalizing constant of the continuous component will turn out to be the observed component of the discrete process which is $\Phi_2(\gamma'x, \alpha'x; \rho_{23})$ in this case. There are various bivariate models that fit into this framework depending on the assumption about the observability of S_1 and S_2. This ranges from separate observability of both S_1 and S_2 to observability of S_1S_2 only (see Meng and Schmidt (1985)).

The extension of this result to more than two-level selection problem is straightforward. For instance, in the three-level selection problem, the continuous component of the sample selection density is a CSN density with dimensions p=1 and q=3. The normalizing constant of this density turns out to be the completely observed part of the discrete component, which is a trivariate probit model with level of observability determined by context.

5.3 Moments and Maximum Likelihood estimator for multilevel selection model

The fact that the continuous component of the multilevel sample selection density is from a well established CSN family results in a straightforward formula for its mean and variance. These models turn out to be generalizations of Heckman's two-step method.

To derive the conditional mean and variance in two-level selection problem, we make use of the mgf of the CSN distribution. The mean is then given by:

$$E(Y|x, S_1^\star > 0, S_2^\star > 0) = \beta'x + \sigma\rho_{12}\Lambda_1(\theta) + \sigma\rho_{13}\Lambda_2(\theta), \tag{5.12}$$

where

$$\Lambda_1(\theta) = \frac{\phi(\gamma'x)\Phi\left(\frac{\alpha'x-\rho_{23}\gamma'x}{\sqrt{1-\rho_{23}^2}}\right)}{\Phi_2(\gamma'x, \alpha'x; \rho_{23})} \quad \text{and} \quad \Lambda_2(\theta) = \frac{\phi(\alpha'x)\Phi\left(\frac{\gamma'x-\rho_{23}\alpha'x}{\sqrt{1-\rho_{23}^2}}\right)}{\Phi_2(\gamma'x, \alpha'x; \rho_{23})}.$$

$\Lambda_1(\theta)$ and $\Lambda_2(\theta)$ are the bivariate inverse Mills ratio. This equation extends Heckman's two-step method (see equation (4.4)) to two-level selection problems. A standard bivariate probit model is fitted depending on what is assumed about the observability of S_1 and S_2 and γ & α are estimated. These are used to construct $\Lambda_1(\hat{\theta})$ and $\Lambda_2(\hat{\theta})$ for cases with S_1 and S_2 greater than zero. These quantities are taken as additional covariates in (5.12) and fitted by least squares. The coefficient of the additional covariates give estimates of $\sigma\rho_{12}$ and $\sigma\rho_{13}$ respectively.

A consistent estimate of the variance can be derived from the conditional

variance given by:

$$\begin{aligned}\operatorname{var}(Y|x, S_1^\star > 0, S_2^\star > 0) =& \sigma^2 - \sigma^2\rho_{12}^2(\gamma' x)\Lambda_1(\theta) - \sigma^2\rho_{13}^2(\alpha' x)\Lambda_2(\theta) \\ &+ \frac{\phi_2(\gamma' x, \alpha' x; \rho_{23})}{\Phi_2(\gamma' x, \alpha' x; \rho_{23})}\Big[2\sigma\rho_{12}\sigma\rho_{13} - \rho_{23}(\sigma^2\rho_{12}^2 + \sigma^2\rho_{13}^2)\Big] \\ &- \Big(\sigma\rho_{12}\Lambda_1(\theta) + \sigma\rho_{13}\Lambda_2(\theta)\Big)^2 \\ =& \sigma^2 + \upsilon.\end{aligned} \tag{5.13}$$

The error terms of the selected sample are heteroscedastic. A generalization of Heckman's estimator for σ^2 given by

$$\sigma^2 = (S - \sum \hat{\upsilon}_i)/N_2,$$

where S is the sum of squared residuals from the second-step regression, N_2 is the size of the complete cases, and υ_i equals $\hat{\upsilon}_i$ after parameter estimates have been substituted for their true values, can be used to get consistent estimator for σ^2.

The derivation of equations (5.12) and (5.13) require evaluation of derivatives of multinormal integrals. Suppose we have a q-dimensional normal random vector $\mathbf{S}$, with mean $\boldsymbol{\nu}$ and a positive definite matrix $\Omega_{q\times q}$ whose elements are $\omega_{i,j}$. The derivative of $\Phi(\mathbf{S}; \boldsymbol{\nu}, \Omega)$ with respect to any S_i is given by (see Dominguez-Molina et al. (2004))

$$\frac{\partial}{\partial \mathbf{s}_i}\Phi_q(\mathbf{s}; \boldsymbol{\nu}, \Omega) = \phi(S_i; \nu_i, \omega_{ii})\Phi_{q-1}\Big(\mathbf{s}_{\neg i}; \boldsymbol{\nu}_{\neg i} + \Omega_{i\neg i}\omega_{ii}^{-1}(S_i - \nu_i), \Omega_{\neg i\neg i} - \omega_{ii}^{-1}\Omega_{i\neg i}\Omega_{i\neg i}'\Big),$$

where $\mathbf{s}_{\neg i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_p)'$, $\Omega_{\neg i\neg i}$ is the $q-1 \times q-1$ matrix derived from $\Omega_{q\times q}$ by eliminating its i-th row and its i-th column and $\Omega_{i\neg i}$ is the $q-1$ vector derived from the i-th column of Ω by removing the i-th row term. The second derivatives is given as

$$\begin{aligned}\frac{\partial^2}{\partial \mathbf{s}_i \partial \mathbf{s}_j}\Phi_q(\mathbf{s}; \boldsymbol{\nu}, \Omega) =& \phi_2(s_{[i,j]}; \boldsymbol{\nu}_{[i,j]}, \Omega_{[i,j]})\Phi_{q-2}\Big(\mathbf{s}_{\neg[i,j]} - \Omega_{[i,j]\neg[i,j]}\Omega_{[i,j]}^{-1}(\mathbf{s}_{[i,j]} - \boldsymbol{\nu}_{[i,j]}); \\ & \boldsymbol{\nu}_{\neg[i,j]}, \Omega_{\neg[i,j]\neg[i,j]} - \Omega_{[i,j]\neg[i,j]}\Omega_{[i,j]}^{-1}\Omega_{[i,j]\neg[i,j]}'\Big),\end{aligned}$$

where the definition is as before but with the components (i, j) taken simultaneously and $\phi_2(.;.,.)$ denotes the PDF of a standard bivariate normal distribution. By convention, $\Phi_0 = 1$.

The log-likelihood function takes the form:

$$
\begin{aligned}
l(\beta,\sigma,\gamma,\alpha,\rho_{12},\rho_{13},\rho_{23}) = \sum_{i=1}^{N}\Bigg(& S_{1i}S_{2i}\Big[\ln f(y_i|x_i, S_{1i}=1, S_{2i}=1)\Big] \\
& + S_{1i}S_{2i}\Big[\ln \Phi_2(\gamma' x_i, \alpha' x_i; \rho_{23})\Big] \\
& + S_{1i}(1-S_{2i})\Big[\ln \Phi_2(\gamma' x_i, -\alpha' x_i; -\rho_{23})\Big] \\
& + (1-S_{1i})S_{2i}\Big[\ln \Phi_2(-\gamma' x_i, \alpha' x_i; -\rho_{23})\Big] \\
& + (1-S_{1i})(1-S_{2i})\Big[\ln \Phi_2(-\gamma' x_i, -\alpha' x_i; \rho_{23})\Big]\Bigg).
\end{aligned} \tag{5.14}
$$

5.3.1 Monte Carlo Simulation

The finite-sample performance of the models in section 5.3 are studied in two parts-the moment based estimator (5.12) and the maximum likelihood estimator (5.14). The outcome equation is $Y_i^\star = 0.5 + 1.5x_i + \varepsilon_{1i}$, where $x_i \overset{iid}{\sim} N(0,1)$ and $i = 1,\ldots,N = 1000$. The two-level selection equations are given as $S_{1i}^\star = 1 + 0.4x_i + 0.3w_i + \varepsilon_{2i}$ and $S_{2i}^\star = 1 + 0.6x_i + 0.7w_i + \varepsilon_{3i}$, where $w_i \overset{iid}{\sim} N(0,1)$. The error terms are generated from a trivariate normal distribution with covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.7 & 0.5 \\ 0.7 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$. This construction implies that the variance of the outcome model is 1.

We only observe values of $Y_i^\star$ when both $S_{1i}^\star$ and $S_{2i}^\star$ are greater than zero. With this representation, roughly 30% of the observations were censored. Roughly 70% censored observations was generated by changing the intercept terms in the selection equations $S_{1i}^\star$ and $S_{2i}^\star$ to -0.1 and -0.2 respectively. In both cases, we allow for full observability in the bivariate process. Pilot simulation results show that there is very little gain in imposing exclusion restriction between the two selection equations, (although this is recommended in practice due to the linearity of the bivariate inverse Mills ratio on a wide range of its support) and as such we did not impose this criteria. Since the moment based method is very common for modeling multilevel sample selection models in practice, we consider four alternative models in this case

- 2TS: Model that generalizes Heckman selection model and accounts for selectivity induced by the selection equations and further impose correlation of the error terms in the selection equations (5.12)

- 2TS0: Model that accounts for selection bias generated by the selection equations, but assumes that the errors in the selection equations are independent
- TS: Classical Heckman two-step method where the selection equations are collapsed to a single indicator for missingness (4.4)
- OLS: Ordinary least square regression using complete cases.

The use of full information maximum likelihood approach to multilevel sample selection problems is not common in the literature. This is due in part to the robustness of the moment based estimator (5.12), to deviation from normality. Nonetheless, we investigate its performance when the underlying normal assumption holds in a simulation study under three model specification. The model labeled 2SNM is the maximum likelihood counterpart of 2TS where correlation is imposed on the error terms of the two selection equations. The SNM1 and SNM2 models are the classical Heckman selection where the two selection equations are collapsed into a single indicator for missingness. Since we have two selection equations, and we do not know the true underlying equation out of the two, the SNM1 model is assumed when the first selection equation is the correct model ($S^{\star}_{1i} = 1 + 0.4x_i + 0.3w_i + \varepsilon_{2i}$) and SNM2 model is assumed for the second selection equation ($S^{\star}_{2i} = 1 + 0.6x_i + 0.7w_i + \varepsilon_{3i}$).

Table 5.1 is the results of the simulation when the likelihood based estimator is used. When interest is not in the selection process, the results shows that collapsing the indicator for missingness and the use of classical Heckman model (SNM1 & SNM2) gives consistent parameter estimates for the outcome as well as the 2SNM model. However, correct specification of the selection model may be difficult in the classical Heckman model since more than one equation now governs the selection process and different covariates might feature in the equations. In addition, it is known that high degree of censoring usually leads to efficiency loss as compared to full data. This however, does not affect the consistency of the parameters as long as the model is correctly specified. In fact, the consistency of model parameters under 70% censored observation, as shown in our simulation result, does not appear worse than the 30% case (although there is increase in the variance of the former). However, results for moment based estimates (see Table 5.2) showed that a high level of censoring might affect the consistency of the estimates. Parameter estimates from OLS are not consistent.

Table 5.1: Simulation results (multiplied by 10,000) for the likelihood based estimator of two-level selection model.

		Bias			MSE		
		2SNM[a]	SNM1[b]	SNM2[c]	2SNM	SNM1	SNM2
$m = 30\%$	β_0	28	62	62	61	44	44
	β_1	-94	7	7	23	21	21
	σ	-37	-115	-115	17	18	18
	γ_0	47	-3717		29	1405	
	γ_1	3	1442		30	235	
	γ_2	21	2456		25	630	
	α_0	38		-3717	33		1405
	α_1	20		-558	32		58
	α_2	56		-1544	36		265
	ρ_{12}	-1352	729		1013	171	
	ρ_{13}	99		1271	253		279
	ρ_{23}	9			32		
$m = 70\%$	β_0	701	544	544	397	309	309
	β_1	-156	-99	-99	50	49	49
	σ	-49	-331	-331	52	57	57
	γ_0	-2	-5212		17	2739	
	γ_1	10	1415		21	226	
	γ_2	90	2359		19	584	
	α_0	-2		-4212	22		1797
	α_1	21		-585	27		60
	α_2	77		-1641	30		297
	ρ_{12}	-1183	-818		917	206	
	ρ_{13}	-38		1182	247		279
	ρ_{23}	1			22		

[a]Maximum likelihood estimator for two-level selection with correlated selection errors.
[b]Heckman selection model with the two-level selection collapsed into the first non-response process.
[c]Heckman selection model with the two-level selection collapsed into the second non-response process.

Table 5.2: Simulation results (multiplied by 10,000) for the moment based estimator of two-level selection model.

		Bias				MSE			
		2TS[a]	2TS0[b]	TS[c]	OLS	2TS	2TS0	TS	OLS
$m = 30\%$	β_0	18	159	212	2842	656	898	74	819
	β_1	-2	-19	-52	-1188	91	121	27	155
	σ	370	4326	-165	-1667	123	5440	26	300
	ρ_{12}	-2305	-3471			6947	10817		
	ρ_{13}	-489	-1999			813	2676		
	ρ_{23}	14	34						
$m = 70\%$	β_0	193	464	742	6857	5943	10811	391	4733
	β_1	-45	-44	-150	-1821	301	540	55	358
	σ	1481	8934	-383	-2711	1051	19577	67	771
	ρ_{12}	-2923	-4247			7079	9836		
	ρ_{13}	-899	-2682			704	2774		
	ρ_{23}	23	22						

[a]Two-step method for two-level selection with correlated selection errors.
[b]Two-step method for two-level selection with independent selection errors.
[c]Classical Heckman two-step method.

Application to the NDI scores

We focus on the measurement at months 8 and use the two-level sample selection model to jointly analyze the two non-response processes in the NDI scores. In line with the study design, 599 patients are expected to return the questionnaire. After removing covariates with missing values, the sample size consists of 567 patients. Out of this, 77 patients returned the questionnaire blank (genuine unit non-response). Vernon (2009) recommended that patients with only 2 missed items should be considered complete, with mean imputation used for adjustment. Rather than discarding these patients, we categorize them as item non-respondents with 43 patients falling into this category. Of course, unit non-respondents are also item non-respondents, making patients with item non-response to be effectively 120. The fully responding units (complete cases) are 447 patients.

The questions to answer are whether unit and item non-response are related and whether both are related to the outcome of interest. To answer the first question, we consider a bivariate probit model with sample selection for unit and item and estimate the correlation parameter. This model is also used to identify possible predictors of non-response in the unit and item equations. Unlike the discrete component of (5.14), the log-likelihood function for a bivariate probit sample selection

model is

$$l(\gamma,\alpha,\rho_{23}) = \sum_{i=1}^{N}\Bigg(S_{1i}S_{2i}\Big[\ln\Phi_2(\gamma' x_i, \alpha' x_i;\rho_{23})\Big] + S_{1i}(1-S_{2i})\Big[\ln\Phi_2(\gamma' x_i, -\alpha' x_i;-\rho_{23})\Big] + (1-S_{1i})\Big[\ln\Phi(-\gamma' x_i)\Big]\Bigg). \tag{5.15}$$

A simulation study (not reported here) showed that if model (5.15) is correctly specified, correct specification includes imposing exclusion restriction on the covariates in the two equations of the unit and item, the model parameters are consistent. In addition, one can test the hypothesis of conditional independence between unit and item non-response using Wald test or the likelihood ratio test. To fit the two step model (5.12) to a two-level selection problem with sample selection between unit and item non-response, the probit model needed in the bivariate inverse Mills ratio is the one given by equation (5.15). This approach was taken by Luca and Peracchi (2006). We consider the maximum likelihood approach to this problem using the NDI scores. Patients may feel that the treatment they received is of no benefit, and thereby discontinue treatment. This will lead to unit non-response rather than item non-response. We therefore include treatment as a possible predictor of unit non-response.

Table 5.3: Probit model for dropout at months 8.

	Missing at 8 months					
	Bivariate Probit			Individual Probit		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value
int(u)[a]	1.085	0.005	0.000	1.019	0.124	0.000
age	0.002	0.000	0.000	0.017	0.005	0.002
sex(f)	0.015	0.006	0.011	0.117	0.138	0.398
physio	0.008	0.006	0.161	0.067	0.134	0.616
int(i)[b]	1.599	0.045	0.000	0.841	0.100	0.000
age	-0.020	0.000	0.000	0.001	0.005	0.914
sex(f)	-0.302	0.008	0.000	-0.062	0.124	0.616
ρ_{23}	0.078	0.147	0.595			

[a]Intercept for unit non-response.
[b]Intercept for item non-response.

The results in Table 5.3 show that there is conditional independence between unit and item non-response for the scores. This was further affirmed by the likelihood

ratio test that compares the maximized values of the log-likelihood in (5.3) with the sum of the log-likelihoods for two simple probit models for unit and item non-response separately.

Table 5.4: Fit of Two-level selection models ($\rho_{23} \neq 0$) & $\rho_{23} = 0$), and Heckman selection model to the NDI scores at 8 months.

	2SNM($\rho_{23} \neq 0$)			2SNM($\rho_{23} = 0$)			SNM[a]		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
				Selection Equations					
int(u)[b]	0.872	0.005	0.000	0.872	0.005	0.000	0.804	0.115	0.000
age	-0.008	0.000	0.000	-0.008	0.000	0.000	0.001	0.004	0.938
sex(f)	-0.129	0.005	0.000	-0.127	0.005	0.000	-0.069	0.125	0.578
physio	0.044	0.005	0.000	0.042	0.005	0.000	0.085	0.122	0.489
int(i)[c]	4.263	0.482	0.000	4.333	0.273	0.000			
age	0.012	0.035	0.745	0.004	0.018	0.830			
sex(f)	1.584	11.032	0.878	1.984	30.090	0.949			
ρ_{23}	0.656	0.810	0.419						
				Outcome Equation					
int	-0.294	0.055	0.000	-0.342	0.061	0.000	-0.260	1.498	0.862
age	0.096	0.001	0.000	0.094	0.001	0.000	0.082	0.027	0.003
sex(f)	0.658	0.031	0.000	0.641	0.031	0.000	0.571	0.722	0.429
physio	-0.354	0.030	0.000	-0.354	0.030	0.000	-0.418	0.716	0.560
base	0.628	0.002	0.000	0.626	0.002	0.000	0.626	0.052	0.000
wad2	-0.072	0.041	0.081	-0.107	0.041	0.009	-0.101	0.976	0.918
wad3	-0.487	0.056	0.000	-0.524	0.056	0.000	-0.517	1.343	0.701
σ	7.453	0.031	0.000	7.377	0.036	0.000	7.388	0.850	0.000
ρ_{12}	-0.500	0.016	0.000	-0.456	0.022	0.000	-0.460	0.503	0.361
ρ_{13}	-0.055	7.554	0.994	0.289	0.767	0.707			

[a]Selection model where unit and item non-response are collapsed into a single indicator for non-response.
[b]Intercept for unit non-response.
[c]Intercept for item non-response.

Table 5.4 contains the results of a two-level sample selection model with $\rho_{23} \neq 0$ & $\rho_{23} = 0$, and the classical full information Heckman sample selection model where a single indicator is used for unit and item non-response. The 'wad' variable stands for Whiplash Associated Disorder (Whiplash describes both the mechanism of injury and the symptoms caused by that injury). It is a categorical variable with grade 3 the most severe neck disability and grade 1 the least before the patient enters the study. The results in the columns with $\rho_{23} \neq 0$ are reported for

completeness sake. This result also strengthen the earlier conclusion about conditional independence of unit and item non-response reported in Table 5.3. Under the model with conditional independence ($\rho_{23} = 0$), separate probit models are used for unit and item missingness for the discrete components of the log-likelihood function given in (5.14). In addition, the classical sample selection model also adduce to the fact that the selectivity generated by unit and item non-response is not different from zero. The classical Heckman model (SNM) also supported the hypothesis of no selection bias (ρ_{12} has p-value = 0.361). The parameter estimates in the outcome equation of the 2SNM($\rho_{23} = 0$) agrees closely with estimates in the SNM model, a further justification that the missingness on the unit and item can be ignored.

5.4 Multilevel extension of the SSNM model

It is also possible to derive a model similar to the SSNM model of chapter 4 in a multilevel selection framework. Suppose we have a joint process where the outcome Y is skewed and the two selection models have skewness parameters zero. The joint distribution can be written in a CSN form. That is,

$$\begin{pmatrix} Y \\ S_1 \\ S_2 \end{pmatrix} \sim CSN_{3,1}\Big\{\boldsymbol{\mu} = (\beta'x, \gamma'x, \alpha'x), \Sigma = \begin{pmatrix} \sigma^2 & \sigma\rho_{12} & \sigma\rho_{13} \\ \sigma\rho_{12} & 1 & \rho_{23} \\ \sigma\rho_{13} & \rho_{23} & 1 \end{pmatrix}, D = (\lambda/\sigma, 0, 0), \nu = 0, \Delta = 1\Big\}.$$

The conditional probability $P(S_1 = 1, S_2 = 1|y, x)$ is

$$\begin{aligned} CSN_{2,1}\Big\{\boldsymbol{\mu} = &\Big[\gamma'x + \rho_{12}\Big(\frac{y - \beta'x}{\sigma}\Big), \alpha'x + \rho_{13}\Big(\frac{y - \beta'x}{\sigma}\Big)\Big]', \Sigma = \Sigma_2^\star, D^\star = (0, 0)', \\ &\nu = \lambda\Big(\frac{y - \beta'x}{\sigma}\Big), \Delta = 1\Big\}, \end{aligned}$$

where $\Sigma_2^\star$ is as defined in section 5.2.2. Since the skewness parameters are zero, we have a normal distribution. This turns out to be the bivariate normal distribution given in equation (5.11). Similarly, the marginal selection process $P(S_1 = 1, S_2 = 1)$ has a bivariate skew-normal distribution

$$SN_2\left\{\begin{pmatrix} \gamma'x \\ \alpha'x \end{pmatrix}, \begin{pmatrix} 1 & \rho_{23} \\ \rho_{23} & 1 \end{pmatrix}, \begin{pmatrix} \frac{-\lambda(\rho_{12} - \rho_{13}\rho_{23})}{(1 - \rho_{23}^2 + \lambda[\rho_{12}^2 + \rho_{13}^2 - 2\rho_{12}\rho_{13}\rho_{23}])} \\ \frac{-\lambda(\rho_{13} - \rho_{12}\rho_{23})}{(1 - \rho_{23}^2 + \lambda[\rho_{12}^2 + \rho_{13}^2 - 2\rho_{12}\rho_{13}\rho_{23}])} \end{pmatrix}\right\}.$$

The continuous component of a multilevel SSNM model has density

$$\frac{\frac{2}{\sigma}\phi\left(\frac{y-\beta' x}{\sigma}\right)\Phi\left(\frac{\lambda(y-\beta' x)}{\sigma}\right)\Phi_2\left(\frac{\gamma' x+\rho_{12}(\frac{y-\beta' x}{\sigma})}{\sqrt{1-\rho_{12}^2}}, \frac{\alpha' x+\rho_{13}(\frac{y-\beta' x}{\sigma})}{\sqrt{1-\rho_{13}^2}}; \frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}}\right)}{P(S_1=1, S_2=1)}. \quad (5.16)$$

The normalizing constant $P(S_1 = 1, S_2 = 1)$ determines the nature of the binary regression model for the discrete process, which is a bivariate binary regression model with skew-normal link. The correlations and the skewness parameter λ contribute to skewness in the model.

It is very unlikely that the likelihood function is tractable in this case. Even in the case where the outcome and the two selection equations are normally distributed, there is possibility of model misspecification and identification issues. Our future work will examine the use of computationally efficient algorithm to estimate parameters in this model and CSN related models.

In principle, one can construct an extension of the CSN distribution with p-dimensional skew-normal and q-dimensional normal random vectors. This model will have all the models we have discussed so far in this thesis as special cases. However, an equivalence of equation 2.18 in the multivariate skew-normal (MSN) distribution (Azzalini and Dalla Valle, 1996) is not readily available. In fact the joint distribution of independent MSN random vectors is not a MSN distribution. A multivariate MSN distribution which satisfies this property is the one proposed by Gupta et al. (2004). One approach to derive the CSN extension is by adding p-dimensional random vector from MSN distribution to an independent q-dimensional random vector from the truncated multivariate normal distribution which involves manipulating complicated algebra.

An alternative approach is to consider a joint distribution of MSN in the CSN form. That is, a $CSN_{p+q,1}$ random vector with p-dimensional MSN and q-dimensional normal components, as demonstrated in the matrix of skewness parameters. Conditioning and marginalization in this representation, and the use of equation (2.17) will result in the required extension. If $p = 1$ and $q = 2$, we have the extension of the SSNM model discussed here. The SSNM model of chapter 4 corresponds to the case with $p = q = 1$.

5.5 Summary

Classical sample selection models and their multilevel counterparts have been in the literature for some time. We have therefore, not claimed any originality in this

proposal. What we have done however, is to unify two streams of literature on this matter and propose a framework for easy generalization to any number of selection equations in a straightforward manner, and which to the best of our knowledge has not been proposed elsewhere.

The econometric literature usually assumes a joint Gaussian error distribution for the outcome and the selection equations. By using properties of truncated normal distribution, the moment-based estimators of sample selection model is derived. On the other hand, the statistics literature contains studies on the closed skew-normal (CSN) distribution. Although the CSN distribution is elegant and a generalization of the Azzalini skew-normal distribution, its use is limited in likelihood based methods due to identifiability issues. When used in sample selection framework, the CSN becomes identifiable due to extra information from the selection process.

We have shown in this thesis that the sample selection models can be constructed either through the use of hidden truncation approach or conditioning in the multivariate normal distribution, and that the latter is a special case of the former in sample selection framework. In addition, it was established that the resulting distribution is the CSN distribution. Using the properties of CSN distribution, moment based estimator for any number of selection equations and with one outcome equation can readily be defined. This gives a unified method for studying more than two-level selection problems which is the current practice in econometric literature. We also emphasize that the density of the sample selection is comprised of a continuous component (CSN) and a discrete component. The model fitted to the discrete component is determined by the marginal distribution of the selection equations. If the marginal distribution is normal, the degree of observability in the discrete process determines the probit model to be fitted and was shown to depend on context.

A simulation study was conducted to assess the performance of the moment and the likelihood based estimators under two-level selection process. Consistent parameter estimates for the outcome models were obtained under the two methods. For the moment based method, the degree of censoring is slightly important. However, the model with 70% censored observations is comparable in terms of precision to the one with 30% censored observation under the likelihood method. In the likelihood method, collapsing the selection process into a single non-response indicator gave less bias in the parameter estimates for the outcome model. The single selection model needs to be correctly specified (a daunting exercise in practice), and there should be no interest in the two selection equations for this to be a reasonable

model. Of course, the results from the classical Heckman model using the collapsed single non-response indicator tends not to work well when the Gaussian assumption is violated.

The NDI scores were analyzed using a multilevel sample selection model in which unit and item non-response (is assumed to) simultaneously affect the outcome of interest. Initial analysis showed that the unit and item non-response are conditionally independent ($\rho_{23} = 0$). A model based on this assumption showed that the dependence between the unit missingness and the outcome model (ρ_{12}), and the dependence between the item missingness and the outcome model (ρ_{13}) are of opposite signs. These offset each other, implying that there may be no selection biases. This was affirmed by using Heckman two-step method, where indicators for the unit and item non-response were collapsed to a single non-response indicator.

On model identifiability, the Fisher information matrix for two selectivity criteria was derived in Ahn (1992) and was shown to be nonsingular. Even in the more than two-level cases, we expect the model to be identifiable. The continuous component (CSN) would necessarily be non-identifiable in general, but will become identifiable from the additional information from the discrete component. However, it is advisable that exclusion restriction is used in the model regardless of the level of observability of the discrete process. The model has better prospects in observational studies and surveys where multilevel selection process need to be analyzed jointly and with information on likely variables that could potentially be responsible for a particular selection process included in the analysis.

Chapter 6

Copula-based sample selection model with sinh-arcsinh distribution as marginals

In chapter 4, we proposed a sample selection model with underlying bivariate skew-normal distribution, the SSNM model. The complexity of the model was reduced by the restriction of the skewness parameter in the selection equation to zero, and maximum likelihood method was used for parameter estimation. Although the skewness parameter of the selection equation was set to zero, the marginal distribution of the selection process is still a univariate skew-normal distribution. We also noted that the correlation parameter ρ in the underlying bivariate process is not adequate to capture association between the outcome and the selection process because of its non-elliptical nature. In particular, the profile likelihood of the skewness parameter, λ has stationarity at $\lambda = 0$.

To circumvent these problems, we present in this chapter the use of copulas in sample selection settings by first showing that the principle of skew distributions arising from selection given in section 2.2, and in particular, equation (2.17) is also the basis of all copula-based sample selection models. Since copulas allow arbitrary marginals, we allow the marginal distribution for the outcome model to follow an asymmetric subfamily of the sinh-arcsinh distribution proposed by Jones and Pewsey (2009), and the selection process to be normally distributed. This model has the advantages of tractability, non-stationarity of profile likelihood if the skewness parameter equals zero and non-singularity of the Fisher information matrix for any parameter value in the model. A simulation study is used to study the finite sample performances of the copula-based models. We also investigate the power of

the Wald and LRT of the hypothesis of symmetry. Motivated by the NDI scores, we assess the ceiling and floor effects of the bounds on the skewness in the data using truncated skew distributions in a sample selection framework. We conclude the chapter by constructing a multilevel selection model using a trivariate Gaussian copula with arbitrary marginals and show that the models in chapter 5 are a special case of this model.

6.1 Copula Theory

Copulas have become a popular tool for multivariate modeling in many applied fields where multivariate dependence structure exists and the validity of the usual multivariate normality assumption is suspect. There is fast growing literature in copula theory (see Joe (1997), Nelsen (2006)). Copulas have been applied in a wide range of problems in biomedical studies (Wang and Wells, 2000; Lambert and Vandenhende, 2002; Escarela and Carriere, 2003). In engineering, copulas are used for hydrological modeling and environmental data (Zhang and Singh, 2006; Genest and Favre, 2007). Applications of copula in sample selection models appeared in much econometric literature (Lee, 1983; Prieger, 2002; Smith, 2003; Genius and Strazzera, 2004).

6.1.1 Basic definitions and theorems

A copula is a function $C : [0,1]^p \to [0,1]$ which satisfy the following properties

1. $C(u_1, \ldots, u_{j-1}, 0, u_{j+1}, \ldots, u_p) = 0$ (grounded property);
   $C(1, \ldots, 1, u_j, 1, \ldots, 1) = u_j$ for all $j \in \{1, \ldots, p\}$, $u_j \in [0,1]$;

2. $C(u_1, \ldots, u_p)$ is non-decreasing in each component u_j;

3. For all $u_{11}, \ldots, u_{p1}$, $u_{12}, \ldots, u_{p2} \in [0,1]^p$ with $u_{i1} \leq u_{i2}$ the following rectangle inequality holds
   $\sum_{i_1=1}^{2} \cdots \sum_{i_p=1}^{2} \Big(-1\Big)^{i_1+\ldots+i_p} C(u_{1i_1}, \ldots, u_{pi_p}) \geq 0.$

Properties 1-3 ensures that a copula is the distribution function of a random vector in $\mathbb{R}^p$ with uniform (0,1) marginals. Property 1 is necessary for the existence of the uniform marginal distributions. Properties 2 and 3 are the usual properties expected of a distribution function. If $F_1(x_1), \ldots, F_p(x_p)$ are univariate distribution functions, then $C(F_1(x_1), \ldots, F_p(x_p))$ is a multivariate distribution function with marginals $F_1(x_1), \ldots, F_p(x_p)$ because $U_j = F_j(X_j)$, $j = 1, \ldots, p$, are uniformly

distributed random variables. Although the definition of copula uses standard uniform marginals, arbitrary marginals can be used in general. We present next a theorem which provides an easy way to form multivariate distributions from known marginals.

Theorem 4. *(Sklar) If F is a distribution function on $\mathbb{R}^p$ with one-dimensional marginal distribution functions $F_1(x_1), \ldots, F_p(x_p)$, then there exists a copula C so that*

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p)). \tag{6.1}$$

If F is continuous, then C is unique and is given by

$$C(u_1, \ldots, u_p) = F(F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p)),$$

for $u = (u_1, \ldots, u_p) \in \mathbb{R}^p$, where $F_i^{-1} = \inf\{x : F_i(x) \geq u\}$, $i, \ldots, p$, is the generalized inverse of F_i.

Conversely, if C is a copula on $[0,1]^p$ and $F_1(x_1), \ldots, F_p(x_p)$ are distribution function in $\mathbb{R}$, then the function defined in (6.1) is a distribution function on $\mathbb{R}^p$ with one-dimensional marginal distribution functions $F_1(x_1), \ldots, F_p(x_p)$. Another fundamental property of a copula is boundedness below and above by Frechet lower and upper bounds, defined as

$$\begin{aligned} F_L(x_1, \ldots, x_p) =& \max\Big[\sum_{j=1}^{p} F_j(x_j) - p + 1, 0\Big] \\ F_U(x_1, \ldots, x_p) =& \min\Big[F_1(x_1), \ldots, F_p(x_p)\Big], \end{aligned}$$

for all $x_1, \ldots, x_p \in \bar{\mathbb{R}}^p$, where $\bar{\mathbb{R}}^p = [-\infty, +\infty]$. This definition implies that the upper bound is always a distribution function while the lower bound is a distribution function only in the bivariate case $p = 2$. For $p > 2$, F_L may be a distribution function under some conditions (Joe, 1997).

6.1.2 Joint and Conditional density functions

For general multivariate distribution, the derivative of the distribution results in its density function. Similar approach can be taken to derive the density function of any copula C with continuous and differentiable marginal distribution. Accordingly, the joint density function is the product of the marginal densities and the copula

density, i.e.

$$f(x_1, \ldots, x_p) = f_1(x_1) \ldots f_p(x_p).c(F_1(x_1), \ldots, F_p(x_p)),$$

where $f_i(x_i)$ is the density corresponding to F_i and c is the copula density, which is defined as

$$c = \frac{\partial^p C}{\partial F_1(x_1), \ldots, \partial F_p(x_p)},$$

(see Kaarik and Kaarik (2009)). The distribution function can also be written in terms of the density as

$$\begin{aligned} C(u_1, \ldots, u_p) =& P(U_1 \leq u_1, \ldots, U_p \leq u_p) \\ =& \int_0^{u_1} \ldots \int_0^{u_p} c(s_1, \ldots, s_p)\, \mathrm{ds_1}, \ldots, \mathrm{ds_p}. \end{aligned} \tag{6.2}$$

If a copula is not absolutely continuous, the joint density does not exist. For the purpose of our work, the idea of conditional distribution is essential. The conditional density of copula C can be easily defined if we take into account the joint density defined earlier and basic definition of conditional density, and is given as follows:

$$\begin{aligned} f(x_p|x_1, \ldots, x_{p-1}) =& \frac{f(x_1, \ldots, x_p)}{f(x_1, \ldots, x_{p-1})} \\ =& \frac{f_1(x_1) \ldots f_p(x_p).c(F_1(x_1), \ldots, F_p(x_p))}{f_1(x_1) \ldots f_{p-1}(x_{p-1}).c(F_1(x_1), \ldots, F_{p-1}(x_{p-1}))} \\ =& f_p(x_p) \frac{c(F_1(x_1), \ldots, F_p(x_p))}{c(F_1(x_1), \ldots, F_{p-1}(x_{p-1}))}, \end{aligned} \tag{6.3}$$

where $c(F_1(x_1), \ldots, F_p(x_p))$ and $c(F_1(x_1), \ldots, F_{p-1}(x_{p-1}))$ are corresponding copula densities.

Aas (2005) categorized copulas into two groups, *implicit* and *explicit* copulas. If the p-dimensional integral in equation (6.2) is implied by well-known multivariate distribution function, we have implicit copulas. For explicit copulas, the p-dimensional integral has a simple closed form. Two examples of implicit bivariate copulas are the Gaussian and Student's t copulas, which are given respectively (in

bivariate form) as

$$C(u_1, u_2; \rho) = \Phi_2\Big(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho\Big) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left\{ \exp\left[-\frac{(x_1^2 - 2\rho x_1 x_2 + x_2^2)}{2(1-\rho^2)} \right] \right\} \mathrm{dx}_1\, \mathrm{dx}_2, \tag{6.4}$$

and

$$C(u_1, u_2; \rho, \eta) = \int_{-\infty}^{t_\eta^{-1}(u_1)} \int_{-\infty}^{t_\eta^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left\{ 1 + \frac{(x_1^2 - 2\rho x_1 x_2 + x_2^2)}{\eta(1-\rho^2)} \right\}^{-(\eta+2)/2} \mathrm{dx}_1\, \mathrm{dx}_2,$$

where $\Phi^{-1}(.)$ is the inverse of the standard univariate normal CDF, t_η^{-1} is the inverse of the standard univariate student-t CDF with η degrees of freedom, expectation 0 and variance $\eta/(\eta-2)$, $\eta > 2$ and ρ $(-1 \leq \rho \leq 1)$ is the Pearson's correlation parameter. These are elliptical copulas, and non-elliptical copulas can be constructed as well.

Clayton copula and Gumbel copula are two examples of explicit copulas, and they belong to the Archimedean family of copula functions. This family has the general form

$$C(u_1, u_2) = \gamma^{-1}\Big(\gamma(u_1) + \gamma(u_2)\Big),$$

where γ^{-1} is the inverse of the strict generator $\gamma(u) : [0,1] \to [0,\infty]$. The dependence parameter δ is embedded in the functional form of the strict generator γ, which is continuous, convex and decreasing function. The unique definition of an Archimedean copula depends on the generator used, which must be a monotone function.

Clayton copula is an asymmetric copula, exhibiting greater dependence in the negative tail than in the positive tail. It is given by the generator $\gamma(u) = \frac{1}{\delta}(u^{-\delta} - 1), 0 < u < 1$, and is of the form

$$C(u_1, u_2; \delta) = \Big(u_1^{-\delta} + u_2^{-\delta} - 1\Big)^{-1/\delta}, \quad \delta \geq 0.$$

Perfect dependence is obtained if $\delta \to \infty$, while $\delta \to 0$ implies independence. The Gumbel copula is also an asymmetric copula, but unlike the Clayton copula, it exhibits greater dependence in the positive tail than in the negative. This copula is

given by

$$C(u_1, u_2; \delta) = \exp\Big[-\Big(-\log u_1^{\delta} - \log u_2^{\delta}\Big)^{1/\delta}\Big], \quad \delta \geq 1,$$

with the generator $(-\log u)^{\delta}$. Perfect dependence is obtained if $\delta \to \infty$, while $\delta = 1$ implies independence. Details about tail dependence of the copulas discussed can be found in Aas (2005). The copula PDFs and h-functions for the bivariate student-t, Clayton and Gumbel copulas are presented in section A.3 in Appendix A.

The simplest copula function is the product copula which corresponds to independence case, and it has the form

$$C(u_1, u_2) = u_1 u_2,$$

but the Gaussian copula is perhaps the most famous copula. Its density takes the form:

$$\begin{aligned} c(u_1, u_2; \rho) =& \phi_2\Big(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho\Big) \prod_{i=1}^{2} \frac{1}{\phi(\Phi^{-1}(u_i))} \\ =& \frac{1}{\sqrt{1-\rho^2}} \exp\Bigg\{-\frac{\rho^2[\Phi^{-1}(u_1)^2 + \Phi^{-1}(u_2)] - 2\rho\Phi^{-1}(u_1)\Phi^{-1}(u_2)}{2(1-\rho^2)}\Bigg\}, \end{aligned}$$

where $\phi_2(.)$ is the PDF of standard bivariate normal distribution. The conditional distribution of the second component given the first in (6.4) is $\partial C(u_1, u_2; \rho)/\partial u_1$, and is the same as the h-function (Aas et al., 2009)

$$h(u_2|u_1; \rho) = \Phi\bigg(\frac{\Phi^{-1}(u_2) - \rho\Phi^{-1}(u_1)}{\sqrt{1-\rho^2}}\bigg). \tag{6.5}$$

The function $h(u_1|u_2; \rho)$ can be equivalently defined. The evaluation of derivatives of multinormal integral can be carried out using the equation given in Dominguez-Molina et al. (2004) for general multivariate normal case, or the use of equation 4 given in Genton et al. (2011) for the standard multivariate normal distribution. Based on this, an equivalent function can be derived for any p-dimensional Gaussian copula. For a trivariate Gaussian copula given by

$$C(u_1, u_2, u_3; \Sigma) = \Phi_3\Big(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3); \Sigma\Big),$$

where

$$\Sigma = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}.$$

The derivative $\partial C(u_1, u_2, u_3; \Sigma)/\partial u_1$ is given as

$$h(u_2, u_3|u_1) = \Phi_2\left[\left(\frac{\Phi^{-1}(u_2) - \rho_{12}\Phi^{-1}(u_1)}{\sqrt{1-\rho_{12}^2}}\right), \left(\frac{\Phi^{-1}(u_3) - \rho_{13}\Phi^{-1}(u_1)}{\sqrt{1-\rho_{13}^2}}\right); \tau_{23|1}\right], \tag{6.6}$$

where $\tau_{23|1}$ is the partial correlation between u_2 and u_3 given u_1, and it is given by

$$\frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}}.$$

Equations 6.5 and 6.6 forms the basis of the models that will be discussed in subsequent sections.

The Gaussian copula is flexible in that it allows for equal degree of positive and negative dependence that includes both Fréchet bounds in its permissible range. However, it is asymptotically independent. This means, regardless of the assumed correlation, extreme tail events appear to be independent in each margins because the density function is thin at the tails. In situations where asymmetric tail dependence is suspected, other measures of dependence such as Kendall's τ and Spearman's ρ can be easily computed from ρ, and like ρ, they take value on [-1,1] which are familiar to applied researchers.

6.2 Sample selection and Gaussian copula

The use of copulas in sample selection framework dated back to the paper by Lee (1983), although he neither mentioned nor gave any reference to its use. He used Gaussian copula, and several authors have extended this ideas to other copulas. Prieger (2002) used a copula approach to model incidence and duration of hospitalisation using the Farlie-Gumbel-Morgenstern system of bivariate distributions (Kotz et al. (2000), section 44.13). Genius and Strazzera (2004) also relax the assumptions of marginal and joint normality. An expository discussion on modeling sample selection using Archimedean copulas was given in Smith (2003). For general references on copulas, see Nelsen (2006).

Arguably, a lot of work has been done on the use of copulas in sample selection settings, none that we are aware of explicitly derive the copula sample selection

model and link it with the general sample selection model. The use of truncated distributions to model bounded scores in sample selection framework is also not adequately studied.

Recall the regression models given in section 3.1, that is

$$Y_i^\star = \beta' x_i + \sigma\varepsilon_{1i}, \quad i = 1, \ldots, N,$$

as regression model of interest, and selection mechanism given as

$$S_i^\star = \gamma' x_i + \varepsilon_{2i}, \quad i = 1, \ldots, N,$$

where all model parameters are as defined in section 3.1. Now we assume that the error distributions have arbitrary marginals and are 'coupled' by a Gaussian copula

$$C(F_1(\varepsilon_{1i}), F_2(\varepsilon_{2i}); \rho) = \Phi_2\Big(\Phi^{-1}(F_1(\varepsilon_{1i})), \Phi^{-1}(F_2(\varepsilon_{2i})); \rho\Big),$$

where F_1 is the error distribution of the outcome margin with corresponding density f_1 and F_2 is the error distribution of the selection process with density f_2. One can easily write

$$F_1(\varepsilon_{1i}) = F_1\Big(\frac{Y_i^\star - \beta' x_i}{\sigma}\Big) \text{ and } F_2(\varepsilon_{2i}) = F_2(S_i^\star - \gamma' x_i).$$

Using the general equation (3.1) and properties of Gaussian copula, we have

$$f(y|x, S = 1; \Theta) = \frac{\frac{1}{\sigma} f_1\Big(\frac{y-\beta' x}{\sigma}\Big) \Phi\left\{ \frac{\Phi^{-1}\Big(F_2(\gamma' x)\Big) + \rho\Phi^{-1}\Big(F_1\Big(\frac{y-\beta' x}{\sigma}\Big)\Big)}{\sqrt{1-\rho^2}} \right\}}{F_2(\gamma' x)}, \tag{6.7}$$

where Θ is the parameters from the models F_1 and F_2, and

$$f(y|x) \equiv \frac{1}{\sigma} f_1\Big(\frac{y - \beta' x}{\sigma}\Big),$$

$$P(S^\star > 0|y, x) \equiv \Phi\left\{ \frac{\Phi^{-1}\Big(F_2(\gamma' x)\Big) + \rho\Phi^{-1}\Big(F_1\Big(\frac{y-\beta' x}{\sigma}\Big)\Big)}{\sqrt{1-\rho^2}} \right\}, \text{ derived from the h-function,}$$

$$P(S^\star > 0) \equiv F_2(\gamma' x).$$

Equation (6.7) is the general continuous component of sample selection density from bivariate Gaussian copula with arbitrary marginals. To examine this, suppose $F_1(\varepsilon_{1i})$ and $F_2(\varepsilon_{2i})$ are normally distributed, then (6.7) reduces to (3.2).

The discrete component can be determined from the marginal distribution of the selection process. In this case, the distribution of $F_2(\gamma' x)$ is used. The binary regression is of the general form:

$$P(S = s) = \{F_2(\gamma' x)\}^s \{1 - F_2(\gamma' x)\}^{1-s}.$$

Roughly speaking, the normalizing constant of the continuous density (6.7) will be the observed component of the binary regression. If F_2 is a CDF of normal distribution, the binary regression becomes the usual probit model.

The log-likelihood function is

$$l(\Theta) = \sum_{i=1}^{N} \Bigg\{ S_i \ln f_1\Big(\frac{y_i - \beta' x_i}{\sigma}\Big) + S_i \ln \Phi \left[\frac{\Phi^{-1}\Big(F_2(\gamma' x)\Big) + \rho \Phi^{-1}\Big(F_1\Big(\frac{y-\beta' x}{\sigma}\Big)\Big)}{\sqrt{1-\rho^2}} \right]$$
$$- S_i \ln \sigma + (1 - S_i) \ln\Big(1 - F_2(\gamma' x_i)\Big) \Bigg\}.$$

Two-step estimation methods have been proposed in copula literature in lieu of the MLE estimation approach. Joe (1997) proposes the Inference Functions for Margins (IFM) which involves maximizing the likelihoods of the marginal models separately. The estimated margins are then combined into a multivariate model to estimate the association parameter. A reverse of the IFM method called Canonical Maximum Likelihood (CML) was proposed by Genest et al. (1995). In the CML method, empirical distribution functions of the margins are first used to estimate the association parameters, and the parameters of the margins are subsequently estimated. However, neither the IFM and CML are appropriate in sample selection settings because the model fitted to the outcome utilizes only selected population in that margin, and thus introduce selection bias.

There are competing skew-normal distributions that can be used as F_1 for the outcome model. The distribution that readily comes to mind is the Azzalini (1985) (or Azzalini-type) SN distribution generated from hidden truncation process by perturbation of the normal kernel. Applied statisticians often look at the SN distribution as the panacea for modeling continuous non-normal data. A possible reason for this is because simple and common nonlinear operations such as truncation, conditioning and censoring carried out on normal random variables lead invariably to versions of skew-normal random variables. The SN distribution therefore appears to be an appropriate model for modeling hidden truncation.

Although researchers are more familiar with the Azzalini SN model, one cannot be sure if there is any hidden truncation present in the underlying process. Body

Mass Index (BMI) data is necessarily skew, but not because of hidden truncation. Even in data where hidden truncation is suspected, the goal has always been to obtain a good fit to the data but not to model the hidden truncation process itself (see Arnold et al. (1993)). Since this is usually the goal of applied researchers, the choice of skew-normal model used may be unimportant as long as the model provides adequate fit to the data and inference is not hampered. Further details on the Azzalini-type SN distributions can be found in chapter 2. We describe next the sin-arcsinh (SHASH) distribution, which is the proposed marginal distribution for the outcome model in this chapter. The SN models are used for comparison purposes and to form links with earlier chapters.

6.3 Sinh-Arcsinh distribution (SHASH)

Several problems are associated with Maximum Likelihood (ML) estimation for the SN distribution. The three well known are:

1. multiple maxima on the likelihood surface (Pewsey, 2000)
2. a solution to the score equations always exists associated with $\lambda = 0$ (Azzalini, 1985; Arnold et al., 1993)
3. the expected information matrix is singular when $\lambda = 0$ (Azzalini, 1985).

The centered parametrization was used to circumvent the last of these problems. However, no satisfactory solution has has been found for the second problem. In fact Pewsey (2006) showed that the second problem is not peculiar to SN distribution but for any skew distribution generated by the perturbation of the normal kernel. Instead of the use of the Azzalini-type SN distribution with their associated inferential problems, other class of skew distribution can be used.

Recently, Jones and Pewsey (2009) proposed the sinh-arcsinh transformation as a general means of generating classes of distributions containing symmetric as well as asymmetric cases with varying tailweight.

Definition 9. *A random variable $Y_{\epsilon,\delta}$, is said to have a sinh-arcsinh normal distribution if its PDF can be written as*

$$f_{\epsilon,\delta}(y) = \frac{1}{\sqrt{2\pi(1+y^2)}}\delta C_{\epsilon,\delta}(y)\exp\{-S^2_{\epsilon,\delta}(y)/2\},$$

where $Z = S_{\epsilon,\delta} \equiv sinh\Big(\delta sinh^{-1}(y) - \epsilon\Big)$ *and* $C_{\epsilon,\delta}(y) = cosh\Big(\delta sinh^{-1}(y) - \epsilon\Big) = \{1 + S^2_{\epsilon,\delta}\}^{1/2}$.

Z is the sinh-arcsinh transformation, ϵ is the skewness parameter with $\epsilon > 0$ corresponding to positive skewness, δ measures the tailweight with $\delta < 1$ yielding heavier tails than the normal distribution and $\delta > 1$ yielding lighter tails. It is easy to see that $f_{0,1}(y) = \phi(y)$, is a standard normal distribution. We will keep the normal tailweight $\delta = 1$ and focus on the skewness parameter ϵ (i.e. $f_{\epsilon,1}(y)$). This is the asymmetric subfamily of the SHASH distribution used in Rosco et al. (2011) to generate skew-t distribution. We still refer to this as the SHASH distribution in what follows. The corresponding CDF is written as $F_{\epsilon,\delta}(y) = \Phi\{S_{\epsilon,1}(y)\}$, where $S_{\epsilon,1}(y) = cosh(\epsilon)y - sinh(\epsilon)(1 + y^2)^{1/2}$. This is used to scale the SHASH PDF in order to derive the truncated version of this distribution as was explained in section 3.4.1.

Unlike the SN density where the introduction of skewness parameter changes the weight in one of the tails, the SHASH density retains two-normal like tails when skewness parameter is introduced. The scale-location extension is of the form $\eta^{-1} f_{\epsilon,1}\{\eta^{-1}(y - \xi)\}$. Analogous to the SN distribution, the parameter ξ is not the mean of the distribution but a function of it. For further details on the SHASH distribution, we refer the reader to Jones and Pewsey (2009).

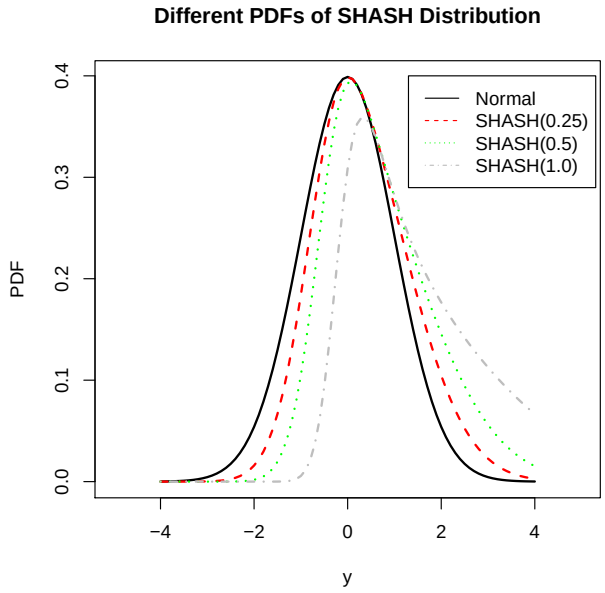

Figure 6.1: Comparison of SHASH densities.

Figure 6.1 shows the densities corresponding to 4 different positive skewness. The plot reinforces the skewness parameter ϵ satisfying the skewness ordering

mention in Jones and Pewsey (2009) for fixed δ. We emphasis that the SHASH and the SN models are different, although it may be somewhat possible to relate the magnitude of the skewness parameter ϵ, for the SHASH model and λ, for the SN model using Arnold and Groeneveld (1995). Figure 6.4 shows the q-q plots of SHASH($\epsilon = 1.0$) and SN($\lambda = 1.0$) margins from a bivariate Gaussian copula with correlation 0.5 and normal margins. The degree of deviation from normality is more pronounced for $\epsilon = 1$ than $\lambda = 1$. This conclusion is supported by the contour plots in Figures 6.2 and 6.3.

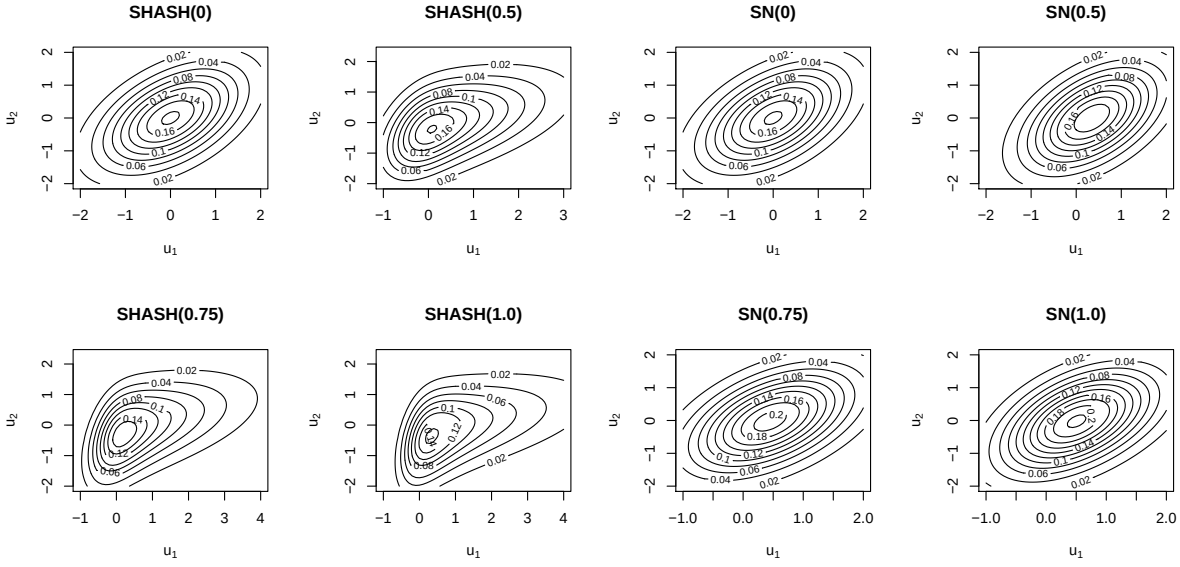

Figure 6.2: Contour plots of SHASH distribution with $\rho = 0.5$ between marginals.

Figure 6.3: Contour plots of SN distribution with $\rho = 0.5$ between marginals.

6.3.1 Monte Carlo Simulation

In this section we study finite sample properties of the MLEs for SHASH and SN copula based sample selection models. We generated the data in the same way as the simulation scenarios given in chapters 3 and 4.

Exploration of SHASH model using Profile likelihoods

We first explore the likelihood surface of the parameters in the SHASH model using profile likelihood on artificially generated data. This will help to study uncertainty in maximum likelihood estimates. To make a fair comparison between the SHASH and

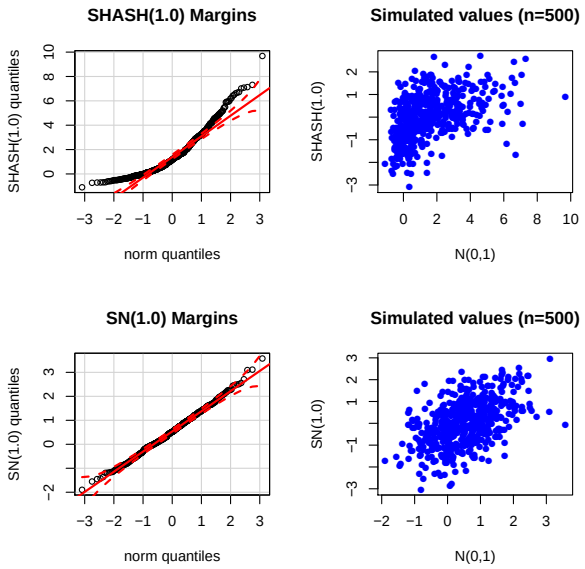

Figure 6.4: Q-Q plots of SHASH($\epsilon = 1.0$) and SN($\lambda = 1.0$) margins from a bivariate Gaussian copula with correlation 0.5 and normal margins.

the SN models, the errors were generated from a bivariate normal distribution. We make use of the selection equation $S_i^\star = 1 + x_i + 1.5w_i + \varepsilon_{2i}$ with exclusion restriction (recall that the outcome model is $Y_i^\star = 0.5 + 1.5x_i + \varepsilon_{1i}$, correlation $\rho = 0.5$, and standard deviation $\sigma = 1$). In summary, the model parameters for data generation are $\Theta = (\beta' = (0.5, 1), \gamma' = (1, 1, 1.5), \sigma = 1, \rho = 0.5,$ and $\epsilon = \lambda = 0)$.

Table 6.1 shows the results of fitting the SHASH, SN, and the correct model, selection normal model (SNM) to the generated data. The three models gave a good fit to the data although the the skewness parameter λ for the SN model is poorly estimated. The SHASH model, according to the log-likelihood value (-1273.72), is the best fitting model (by a very small margin). It fits better than the correct SNM model that generated the data. The Wald test for the hypothesis of symmetry agrees closely with the data generation process with the skewness parameters in the SHASH and the SN models nonsignificant at 5% level of significance. A LRT for symmetry between SHASH and the SNM model also support this conclusion (p-value =0.842). Note that the LRT cannot be carried out between the SN and the SNM model.

Figures 6.5-6.8 are the profile likelihoods for the parameters in the SHASH, SN and SNM models. The profile likelihood for λ is very flat in the neighborhood of zero. An approximate 95% likelihood ratio confidence interval is wide (-1.25,

Table 6.1: Fit of SHASH model, SN model, and classical Heckman model (SNM) to a sample selection dataset with bivariate normal error distribution.

	SHASH			SN			SNM		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
				Selection Equation					
int	0.999	0.073	0.000	1.003	0.073	0.000	0.999	0.073	0.000
x	1.093	0.080	0.000	1.097	0.080	0.000	1.091	0.080	0.000
w	1.453	0.094	0.000	1.461	0.094	0.000	1.455	0.090	0.000
				Outcome Equation					
int	0.545	0.093	0.000	0.277	0.669	0.679	0.538	0.053	0.000
x	1.520	0.045	0.000	1.526	0.045	0.000	1.525	0.045	0.000
σ	1.033	0.029	0.000	1.063	0.166	0.000	1.030	0.029	0.000
ρ	0.336	0.107	0.002	0.334	0.106	0.002	0.330	0.106	0.002
ϵ & λ	-0.009	0.055	0.870	0.324	0.859	0.706	-	-	-
Loglik	-1273.72			-1273.76			-1273.74		

1.07). Not only is the interval very wide, the likelihood surface possibly has multiple maxima as well. The profile likelihood for λ shows that it attains maximum with the value -1273.71, and three values of λ (-0.450, -0.475, -0.500) correspond to this value. It should be noted that the SN model is sensitive to initial values, but this sensitivity appears to only affect the estimation of λ. For the SHASH model however, the bias in the estimation of ϵ, its skewness parameter, is small. The likelihood surface is very steep and its interval is rather precise (-0.12,0.07) with corresponding maximum likelihood -1273.70. The corresponding ϵ, is -0.025, which is not far from the estimated -0.009 in the full model.

Monte Carlo study

We carry out a full simulation study with 1000 replications using the SHASH, SN, SNM models, and the Heckman two-step method (TS). The error terms are generated from a bivariate Gaussian copula with SHASH distribution as the marginal distribution for the outcome with ϵ= 0, 0.25, 0.5 and 1.0. We include simulation results from the SN models for completeness sake only as the SHASH and the SN models are not comparable except when $\epsilon = 0$. Table 6.2 shows the results of the simulation under these models in the presence of the exclusion restriction. The bias in the estimates of skewness parameters when $\epsilon = 0$ is lower in the SHASH model than the SN, which support our earlier observation in Table 6.1. The outcome mod-

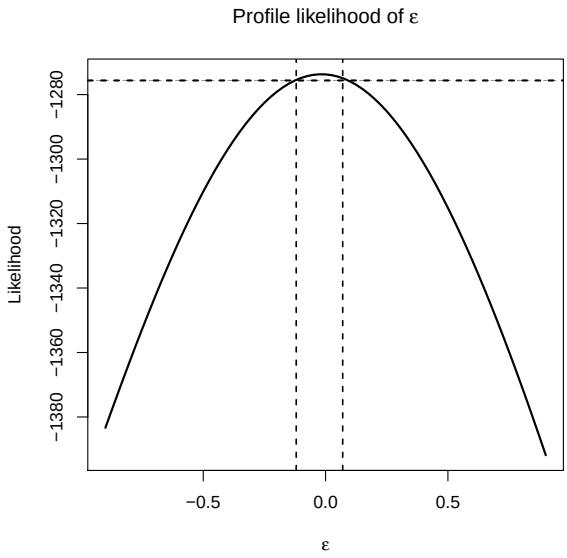

Figure 6.5: Profile likelihood for ϵ using SHASH model. Data generated from a bivariate normal distribution with $\rho = 0.5$.

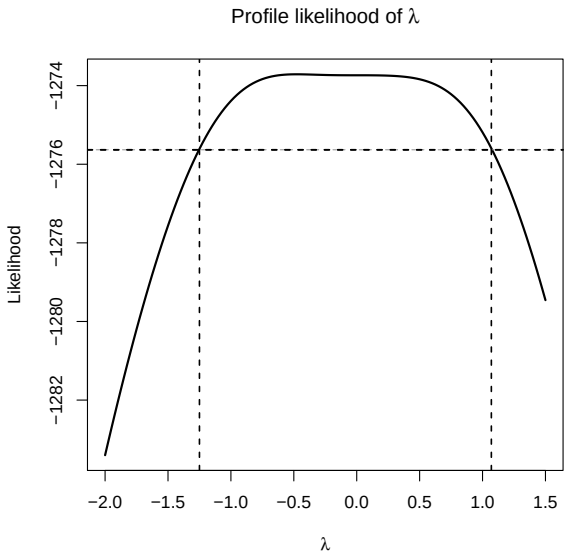

Figure 6.6: Profile likelihood for λ using SN model. Data generated from a bivariate normal distribution with $\rho = 0.5$.

els are generally better estimated under the SHASH model than any of the other models. However, the estimate of the selection part of the model is poor. Similar observations can be seen in Table 6.3 in the absence of exclusion restriction.

Test of hypothesis of symmetry in SHASH models

We investigate the performance of two tests for testing hypothesis of symmetry in the SHASH model when used as an outcome model in a bivariate Gaussian copula with normal selection process. The tests under consideration are the Wald test of the hypothesis $H_0 : \epsilon = 0$ and the LRT of symmetry which has χ^2_1 distribution. The data is generated as before but we restrict attention to the exclusion restriction scenario. We consider varying sample sizes ($N = 500, 1000$) and varying correlation ($\rho = 0.0$, 0.1, 0.3, 0.5 and 0.7). The nominal level used is 0.05.

When the errors are uncorrelated and $N = 1000$, the LRT maintains correct nominal value. The performances of the Wald test are poorer when the correlation is above 0.5 for $N = 500$ and $N = 1000$. In these cases, the LRT maintains nominal value, especially when $N = 1000$ (see Table 6.4). Tables 6.5 contains powers of the test of hypothesis of symmetry using Wald and LRT. The LRT is a powerful test for symmetry for large sample size ($N = 1000$) and the skewness parameter $\epsilon > 0.2$.

Table 6.2: Simulation results (multiplied by 10,000) in the presence of exclusion restriction.

		Bias				MSE			
		SHASH	SN	SNM	TS	SHASH	SN	SNM	TS
$\epsilon = 0.0$	β_0	-16	-100	-1	2	67	31	24	28
	β_1	26	23	-3	-5	18	18	18	19
	γ_0	85	75	67	73	57	56	50	51
	γ_1	79	67	52	59	59	58	59	60
	γ_2	148	134	98	106	99	98	93	9
	σ	-23	-7	-9	-7	9	9	9	9
	ρ	40	33	-6	-21	83	83	84	113
	ϵ	-12	87	-	-	26	13	-	-
$\epsilon = 0.25$	β_0	21	-5936	3296	3381	84	3819	1115	1174
	β_1	64	-28	53	10	21	23	22	22
	γ_0	212	-152	39	76	67	73	51	53
	γ_1	215	-201	14	85	71	80	62	65
	γ_2	309	-183	10	138	118	134	99	101
	σ	-10	3959	553	535	10	1683	41	39
	ρ	108	-118	242	-5	87	117	114	131
	ϵ	-33	-	-	-	35	-	-	-
$\epsilon = 0.5$	β_0	38	-5819	6779	6972	96	3597	4636	4899
	β_1	53	-80	119	21	21	38	31	28
	γ_0	298	20	5	76	90	141	51	53
	γ_1	300	1	-57	85	94	155	64	65
	γ_2	436	54	-126	138	162	218	109	101
	σ	-24	7182	1870	1825	15	5287	365	348
	ρ	49	-264	451	-38	98	207	152	140
	ϵ	-18	-	-	-	44	-	-	-
$\epsilon = 1.0$	β_0	333	-2992	15238	15720	265	1054	3320	4790
	β_1	71	-137	294	69	31	56	78	57
	γ_0	709	-121	-54	76	267	151	52	53
	γ_1	695	-216	-172	85	267	175	71	65
	γ_2	103	-205	-346	138	486	295	141	101
	σ	131	14997	7038	6919	99	22684	4994	4824
	ρ	-32	-979	704	-126	153	394	224	152
	ϵ	-157	-	-	-	185	-	-	-

Table 6.3: Simulation results (multiplied by 10,000) in the absence of exclusion restriction.

		Bias					MSE		
		SHASH	SN	SNM	TS	SSNM	SN	SNM	TS
$\epsilon = 0.0$	β_0	197	103	154	49	105	84	84	124
	β_1	-117	-124	-121	36	67	63	62	89
	γ_0	61	61	66	66	35	35	38	38
	γ_1	52	54	100	101	47	47	52	52
	σ	-19	-19	-18	59	13	13	12	23
	ρ	-403	-414	-427	-237	500	466	452	651
	ϵ	-30	71	-	-	29	1	-	-
$\epsilon = 0.25$	β_0	134	-5906	2606	3365	130	3886	809	1279
	β_1	-3	-315	651	15	76	107	142	97
	γ_0	73	17	-140	32	37	39	38	36
	γ_1	62	-10	-329	57	50	53	69	49
	σ	46	4283	982	666	15	1940	123	71
	ρ	-176	-1076	1457	-166	490	803	781	684
	ϵ	-62	-	-	-	34	-	-	-
$\epsilon = 0.5$	β_0	64	-5911	5027	6887	148	3962	2728	4926
	β_1	55	-93	1720	77	80	197	459	120
	γ_0	75	-90	-605	32	39	52	76	36
	γ_1	63	-144	-1247	59	51	80	239	49
	σ	67	7629	2935	2017	19	5923	908	443
	ρ	-74	-819	3020	-123	451	1120	1549	681
	ϵ	-56	-	-	-	39	-	-	-
$\epsilon = 1.0$	β_0	939	163	11555	15479	01291	5234	13869	24327
	β_1	723	1786	4066	220	845	1233	2106	242
	γ_0	-92	-843	-1878	35	134	226	417	35
	γ_1	-249	-1511	-3244	63	230	596	1215	49
	σ	1670	14123	9300	7263	1109	20642	8776	5355
	ρ	813	867	3954	-113	1075	1859	2301	679
	ϵ	-1465	-	-	-	915	-	-	-

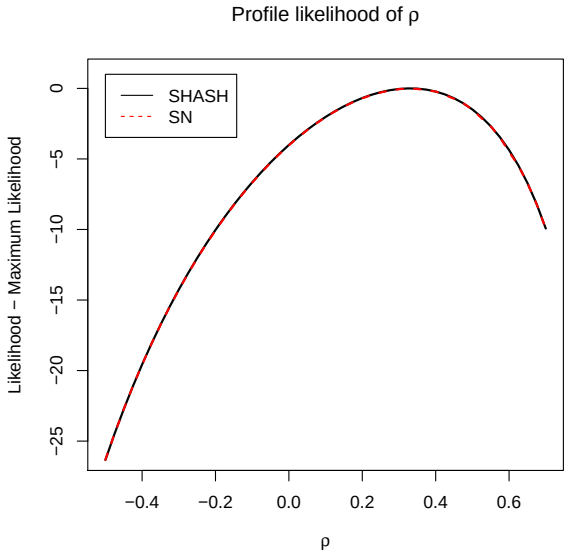

Figure 6.7: Profile likelihood for ρ using SHASH and SN model. Data generated from a bivariate normal distribution with $\rho = 0.5$.

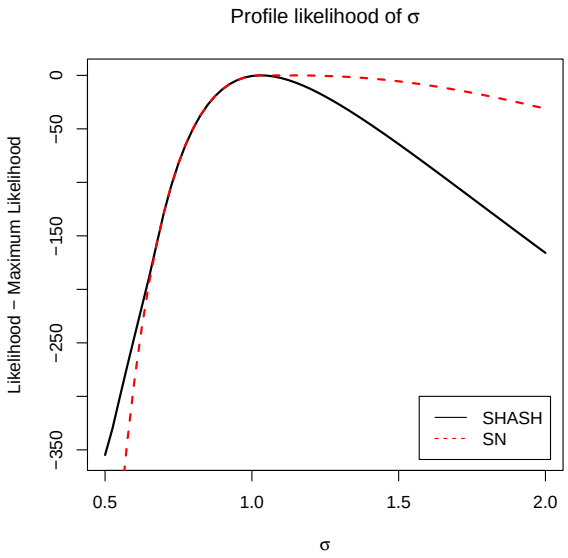

Figure 6.8: Profile likelihood for σ using SHASH and SN model. Data generated from a bivariate normal distribution with $\rho = 0.5$.

Application to the NDI scores

We apply the copula models to the NDI scores at month 8 in this section. Since the outcomes are bounded, it may be of interest to evaluate the effect of the boundedness on parameter estimates, and most importantly, its implications on inherent skewness in the data set. To achieve this, we first consider a case where the outcome model has a truncated distribution for the SHASH and SN models but with normal selection marginals (see section A.4 in Appendix A for coding). The second case is the unrestricted space for the outcomes, that is, the SHASH and the SN models are non-truncated. Table 6.6 contains the results of the truncated SHASH and SN models in the interval [0,50], while Table 6.7 shows the results of the non-truncated models. The two tables gave contradictory conclusions on the importance of skewness in the data. Table 6.6 suggested that the data only appear to be skew because of the bounds, and that the inherent skewness is not important when the bounds in the data are taken into account. However, the skewness parameters in the SHASH and the SN models become significant when the obviously restricted data set is 'forced' on the whole real line. Even in terms of fit, the truncated models have better fits than their respective non-truncated models. For instance, the log-likelihood value for the truncated and untruncated SHASH models are -1424.94 and -1453.21

Table 6.4: Empirical significance levels (as %) of the tests of symmetry for the nominal significance level $\alpha = 0.05$ in the SHASH model.

	$N = 500$		$N = 1000$	
ρ	Wald	LRT	Wald	LRT
0.0	4.7	5.4	4.8	5.0
0.1	5.3	5.2	5.0	4.8
0.3	4.2	5.1	4.8	4.7
0.5	3.6	5.1	4.1	5.0
0.7	3.5	4.7	3.7	5.1

Table 6.5: Powers (as %) of the tests of symmetry for the nominal significance level $\alpha = 0.05$ in the SHASH model.

		$N = 500$		$N = 1000$	
	ρ	Wald	LRT	Wald	LRT
$\epsilon = 0.1$	0.0	25.2	25.7	45.6	44.7
	0.1	24.6	24.5	44.6	44.3
	0.3	24.0	24.8	43.6	42.9
	0.5	24.1	24.9	42.8	42.7
	0.7	24.7	25.0	42.6	42.1
$\epsilon = 0.2$	0.0	71.6	72.3	94.8	95.5
	0.1	72.0	72.7	95.5	95.4
	0.3	71.0	71.3	95.2	95.1
	0.5	69.4	69.6	94.5	94.7
	0.7	69.9	69.4	94.4	94.4
$\epsilon = 0.25$	0.0	88.3	87.9	99.4	99.5
	0.1	86.7	87.4	99.5	99.6
	0.3	86.3	86.8	99.2	99.4
	0.5	86.7	86.5	99.1	99.6
	0.7	86.5	86.8	99.1	99.4

respectively. As expected, parameter estimates for the SHASH and SNM model in Table 6.6 are similar since the skewness parameter is not different from zero. This is not the case in Table 6.7.

A comparison of the SN model in Table 6.7 with the SSNM model in Table 4.4 of chapter 4 shows that the parameter estimates for the outcome models are similar. The estimate of skewness parameter λ, in the models are 1.552 and 1.537 respectively, and both are significant. The treatment effect is not significant in both models. Notice that the correlation ρ is not significant in Table 6.7 but significant in Table 4.4 under the Wald test (although LRT is not significant, p-value = 0.437). Both SN and the SSNM models are possibly misspecified by not taking into account the bounds in the data.

Failure in taking into account bounds in a data set is not the only problem. Ignoring the effect of selection process in a model when it is present can also inflate type 1 error. Consider the truncated skew-normal (TSN) model of chapter 3. The truncation points are taken into account, yet skewness is present in the data. The truncation points were also taken into account in Table 6.6 but with additional information from the selection process. The skewness parameter λ is no longer significant in this case. This further emphasis the relationship between selection and skewness and the importance of dealing with them simultaneously when they are suspected to be present in a data set.

Table 6.6: Fit of copula-based Sinh-archsinh (SHASH), Skew-normal (SN), and Selection-normal model (SNM) sample selection models to the NDI scores at 8 months. The corresponding outcome models are truncated at [0,50].

	SHASH			SN			SNM		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
				Selection Equation					
int	0.834	0.102	0.000	0.827	0.103	0.000	0.829	0.103	0.000
age	0.022	0.006	0.000	0.023	0.006	0.000	0.023	0.006	0.000
sex(f)	0.336	0.131	0.011	0.351	0.134	0.009	0.347	0.133	0.009
				Outcome Equation					
int	-3.857	1.282	0.003	-7.037	2.052	0.001	-3.891	1.293	0.003
age	0.109	0.031	0.001	0.105	0.031	0.001	0.107	0.031	0.001
prev	0.889	0.055	0.000	0.864	0.052	0.000	0.872	0.051	0.000
physio	1.370	0.757	0.071	1.219	0.738	0.099	1.274	0.738	0.085
σ	7.227	0.603	0.000	7.597	0.854	0.000	6.904	0.404	0.000
ρ	0.769	0.082	0.000	0.719	0.010	0.000	0.737	0.086	0.000
ϵ & λ	-0.078	0.099	0.433	0.693	0.518	0.182	-	-	-
Loglik	-1424.94			-1425.11			-1425.26		

6.4 Multilevel Sample Selection

Recall the models

$$Y_i^\star = \beta' x_i + \sigma\varepsilon_{1i}, \quad i = 1, \ldots, N,$$

Table 6.7: Fit of copula-based Sinh-archsinh (SHASH), Skew-normal (SN), and Selection-normal model (SNM) sample selection models to the NDI scores at 8 months. The corresponding outcome models are untruncated.

	SHASH			SN			SNM		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
				Selection Equation					
int	0.828	0.115	0.000	0.822	0.109	0.000	0.835	0.100	0.000
age	0.024	0.006	0.000	0.025	0.006	0.000	0.024	0.006	0.000
sex(f)	0.363	0.156	0.020	0.379	0.144	0.009	0.335	0.129	0.010
				Outcome Equation					
int	1.261	0.852	0.140	-3.553	0.896	0.000	0.799	0.621	0.199
age	0.031	0.031	0.320	0.069	0.032	0.034	0.086	0.023	0.000
prev	0.670	0.035	0.000	0.678	0.035	0.000	0.687	0.035	0.000
physio	0.717	0.528	0.175	0.766	0.534	0.152	0.887	0.538	0.100
σ	5.450	0.453	0.000	7.621	0.548	0.000	6.166	0.292	0.000
ρ	-0.288	0.591	0.626	0.528	0.448	0.239	0.794	0.076	0.000
ϵ & λ	0.236	0.082	0.004	1.552	0.517	0.003	-	-	-
Loglik	-1453.21			-1452.87			-1455.03		

as regression model of interest, and selection mechanisms given as

$$\begin{aligned} S_i^\star &= \gamma' x_i + \varepsilon_{2i}, \quad i = 1, \ldots, N, \\ S_{2i}^\star &= \alpha' x_i + \varepsilon_{3i}, \quad i = 1, \ldots, N, \end{aligned}$$

where $S_{1i} = I(S_{1i}^\star > 0)$, $S_{2i} = I(S_{2i}^\star > 0)$ and $Y_i = Y_i^\star S_{1i} S_{2i}$. Suppose the error terms are arbitrary and can be 'coupled' using a trivariate Gaussian copula. That is,

$$C(F_1(\varepsilon_{1i}), F_2(\varepsilon_{2i}, F_3(\varepsilon_{3i}); \Sigma) = \Phi_3\Big(\Phi^{-1}(F_1(\varepsilon_{1i})), \Phi^{-1}(F_2(\varepsilon_{2i})), \Phi^{-1}(F_3(\varepsilon_{3i})); \Sigma\Big),$$

where Σ is as defined in section 6.1.2 and F_1 is the error distribution of the outcome margin with corresponding density f_1, and F_2 and F_3 are the error distributions of the selection processes with densities f_2 and f_3. Using the link between sample selection and skew distribution, one can write

$$f(y|x, S_1 = 1, S_2 = 1; \Theta) = \frac{\frac{1}{\sigma} f_1\Big(\frac{y - \beta' x}{\sigma}\Big) \Phi_2\Big(A, B; \tau_{23|1}\Big)}{C\Big\{F_2(\gamma' x), F_3(\alpha' x); \rho_{23}\Big\}}, \tag{6.8}$$

where $\tau_{23|1}$ is as defined in section 6.1.2, and

$$A = \left\{ \frac{\Phi^{-1}\Big(F_2(\gamma' x)\Big) + \rho_{12}\Phi^{-1}\Big(F_1\Big(\frac{y-\beta' x}{\sigma}\Big)\Big)}{\sqrt{1-\rho_{12}^2}} \right\}, \; B = \left\{ \frac{\Phi^{-1}\Big(F_3(\alpha' x)\Big) + \rho_{13}\Phi^{-1}\Big(F_1\Big(\frac{y-\beta' x}{\sigma}\Big)\Big)}{\sqrt{1-\rho_{13}^2}} \right\}.$$

Equation (6.8) can be extended readily to more than two selection equations. The model for selection is governed by the bivariate Gaussian copula $C\Big\{F_2(\gamma' x), F_3(\alpha' x); \rho_{23}\Big\}$. In particular, if the error distribution in (6.4) and (6.4) follows trivariate Gaussian distribution then, $F_1(\varepsilon_{1i})$, $F_2(\varepsilon_{2i})$ and $F_3(\varepsilon_{3i})$ are normally distributed marginally, and the model reduces to the model discussed in chapter 5. In this case, the selection model is a bivariate probit model.

The generalization of copula sample selection model to multilevel settings can be carried out for any multivariate copulas that is differentiable. The main issue is to derive the h-function, and for the copulas to have extension beyond the bivariate form.

6.5 Summary

We have shown in this chapter that the link between sample selection and skew distribution can be extended to copula based sample selection models. The copula representation of sample selection models have the advantage that it allows different model specification for the marginals and great flexibility in the specification of the association parameters. This prompted us to consider a flexible class of skew distribution of Jones and Pewsey (2009). Since the focus of the thesis is on modeling skewness in data sets subjected to selective reporting, we have focused on the asymmetric subfamily of this distribution, which we referred to as SHASH model. This model was used as the marginal model for the outcomes. We assumed normal distribution for the margin of selection process throughout. This margins are assumed to be 'coupled' by the bivariate Gaussian copula.

We remark that the choice of a bivariate Gaussian copula was motivated by its flexibility in that it allows for equal positive and negative dependence that includes the Fréchet bounds in its permissible range. In fact a measure of association that takes value on [-1,1], the correlation ρ in this case, is what is easily interpretable to applied researchers as a measure of linear association. Of course, other copulas with this range of association, which can also capture asymmetry in real data, can be constructed. In addition, the use of Gaussian copula allows direct comparison between the models proposed in chapters 4 and 5 with the model in this chapter.

In section 6.1.2, we emphasized the importance of the conditional distribution in both explicit and implicit copulas. This distribution turns out to be the h-functions for conditional copulas discuss in Aas et al. (2009). With the h-function, the continuous component of the sample selection density can be easily constructed. The selection process is then determined from the marginal distribution of the selection process. This was illustrated with equation (6.7), and it is very general for any differentiable bivariate copulas with arbitrary margins. The use of Azzalini (1985) (SN) model was also investigated as a plausible outcome model.

An attempt to gain insight into finite sample properties of the MLEs for the SHASH and the SN models was partially successful. Although the data sets were generated using the SHASH model as marginal, yet the selection parts still have some bias. This is why we investigated powers of the tests of hypothesis of symmetry rather than selection bias, even though the latter is usually of interest in sample selection framework. The impact of bounds on skewness was investigated using truncated versions of the SHASH and the SN distributions as models for the outcomes. The result underscores the importance of bounds and adjustments for selection bias on skewness and possible inflated type-1 error.

Part II

Sensitivity Analysis for Recurrent Event Data with Dropout

Chapter 7

Sensitivity Analysis for Recurrent Event Data Trials subject to informative Dropout

The studies that motivated this work seek to analyze processes which generate events repeatedly over time. Such processes are referred to as recurrent event processes. Examples include seizures in epileptic studies, flares in gout studies or occurrence of cancer tumors.

Interest lies in understanding the underlying event occurrence process. This includes the investigation of the rate at which events occur, the inter-individual variation, and most importantly, the relationship between the event occurrence and covariates such as treatment. One considerable challenge in analyzing recurrent event data arises when a large proportion of patients discontinue before the end of the study, e.g. due to adverse events, leading to partially observed data. Any analysis of such data relies on untestable assumptions regarding the post-discontinuation behavior of patients that drop out. Regulatory agencies are therefore increasingly asking for sensitivity analyses which assess the robustness of conclusions across a range of different assumptions. Sophisticated sensitivity analyses for continuous data, e.g. using the pattern-mixture model approach, are being increasingly performed. However, this is less the case for recurrent event or discrete data.

We present in this chapter some approaches for performing sensitivity analyses for recurrent events data, subject to dropouts, using frequentist multiple imputation (MI) techniques. The modeling approach for recurrent event data used is based on event counts and the traditional framework for analysis is the Poisson process. Poisson models are often used in regression analysis of count data. The constraint

of equal mean and variance is generally inapplicable, and individual effects (random effects) are included in the model. A convenient model is the negative binomial model (Lawless, 1987a,b), which we use in this chapter. We investigate the importance of varying event generation process and the impact of the imputation methods used. In particular, we consider an approach for imputation similar to Little and Yau (1996), which involves imputation of values of missing outcomes using a model that conditions on an 'assumed' treatment received by patients after dropout. We refer to this as the use of event rate different from the MAR rates for imputation in treated arm, since MAR assumption requires that the same observed data and covariates (e.g. treatment) will have the same statistical behavior in their future evolution whether they are observed or not. A method that involves imputation of missing values in the active arm using the event rates of the placebo arm is considered, and we referred to this as placebo multiple imputation (pMI). Imputation of missing values in the active arm using incremental event rates for the active arm is also considered. Of course, patients in the placebo arm are imputed under the MAR missingness assumption in both cases because they receive no active substance in the first instance.

7.1 Motivating Example- The Bladder Cancer Trial

The data used in the second part of this thesis is a publicly available placebo controlled trial of tumor recurrence in patients with bladder cancer. The data is from the bladder tumor study conducted by the Veteran Administration Co-operative Urological Research Group (VACURG). This randomized clinical trial, in its original form (see Andrews and Herzberg (1985)), studied the effect of three treatments on the frequency of recurrence of bladder cancer. The data has been used by Dean and Balshaw (1997), Wellner and Zhang (2000), Sun and Wei (2000), Sun and Wei (2002), Zhang (2006), and Balakrishnan and Zhao (2009). There were 116 patients in the study and all had superficial bladder tumors when they entered the trial. The patients were assigned randomly to one of the three treatments: placebo, pyridoxine and thiotepa.

Table 7.1 shows the distribution of the number of recurrences observed for the patients in each of the three groups. The placebo group has 47 patients while pyridoxine and thiotepa groups have 31 and 38 patients respectively. Table 7.2 gives summary statistics for the follow-up times in the three groups. The placebo group has the highest follow-up time of 64 months.

The version of the data we use in this thesis included only the placebo and

Table 7.1: Distribution of the Number of Recurrences observed for the patients the three treatment groups in bladder cancer trial.

	Num. of Recurrences										Num. of
Treatment	0	1	2	3	4	5	6	7	8	9	patients
Placebo	18	10	4	6	2	4	1	0	1	1	47
Pyridoxine	16	5	4	0	0	2	0	0	2	2	31
Thiotepa	20	8	3	2	2	2	0	1	0	0	38

Table 7.2: Summary statistics for the follow-up times of patients in the three treatment groups in bladder cancer trial.

	Follow-up times in Months					
Treatment	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Placebo	1	23.00	30.00	32.51	43.00	64
Pyridoxine	2	12.50	37.00	32.03	45.50	60
Thiotepa	1	18.25	32.50	31.13	44.00	59

thiotepa groups. Other authors have also considered this version (Wei et al., 1989; Metcalfe and Thompson, 2007). The data is readily available in the survival package in R software. The outcome variable was the timing of the clinical visit in which a recurrence was detected. Patients are censored when they die or when the end of the study is reached. Two baseline variables, the number and size of the tumors removed prior to recruitment to the study are included in the data. Patients in the placebo arm may likely experience higher tumor recurrence rates than patients in the thiotepa group, making it more likely for them to withdraw early. It is therefore very important to handle missing data carefully in recurrent event data settings. Concepts of missing data in recurrent event framework are the same as the ones given in section 2.5.

7.2 Notation and Models

Our study deals with dropouts in recurrent event data, and as such we adopt the same notation as was used in Akacha and Benda (2010).

7.2.1 Notation

Suppose m independent subjects are randomized equally into a two arm trial of an active treatment and placebo, and that each subject experiences a type of recurrent event. Let $N_i(T) = n_i$ denote the number of events over the complete study period [0,T] for the ith subject. The event times j for subject i are denoted by

$0 < t_{i1} < \ldots < t_{in_i} \leq T$ and the corresponding random variables by $T_{i1}, \ldots, T_{in_i}$. Let the treatment be denoted as X_i which is one for the treated group and zero otherwise. Furthermore, let $N_i = \{N_i(T), T_{i1}, \ldots, T_{in_i}\}$ denote the complete recurrent event data information for subject i and let $t_{id} \in (0, T]$ indicate the dropout time for subject i. Then $N_{i,\text{obs}} = \{N_i(t_{id}), T_{i1}, \ldots, T_{id}\}$ denote the observed part and $N_{i,\text{mis}} = \{N_i(T), T_{id+1}, \ldots, T_{in_i}\}$ the missing part of the recurrent event data sequence. Monotone dropout is expected in this setting: when a subject drops out they never return to the study.

7.2.2 Poisson Process Models

A Poisson process is a stochastic process with events occurring randomly over time. Let $N(t)$ be as defined above and let $\lambda(t)$ be a left continuous function such that

$$\int_0^t \lambda(u)\mathrm{d}u = \Lambda(t) < \infty. \tag{7.1}$$

Then, $\{N(t)\}_{t=0}^{\infty}$ is a Poisson process with intensity function $\lambda(t)$ and cumulative intensity function $\Lambda(t)$ if and only if

1. N(0)=0
2. $\Pr\{N(t+h) - N(t) = 0 | H(t)\} = 1 - \lambda(t)h + o(h)$
3. $\Pr\{N(t+h) - N(t) = 1 | H(t)\} = \lambda(t)h + o(h)$
4. $\Pr\{N(t+h) - N(t) > 1 | H(t)\} = o(h)$,

for small h and $t > 0$, and where $o(h)$ is such that $o(h) = \lim_{h\to 0} o(h)/h = 0$. The history of the process, $H(t)$, is the record of all events up to time t, i.e. $H(t) = \{N(u) : 0 \leq u < t\}$. The intensity function ($\lambda(t)$) is the instantaneous probability of an event occurring at a certain time, given the history of the process. Mathematically,

$$\lambda(t) = \lim_{h\to 0} \frac{\Pr(\Delta N(t) = 1 | H(t))}{h},$$

where $\Delta N(t) = N(t+h) - N(t)$ denote the number of events in the interval $[t, t+h)$. If $\lambda(t) = \lambda$, then the Poisson process is said to be a homogeneous Poisson process. A non-homogeneous Poisson process has an intensity function that is time dependent. For Poison processes (homogeneous or non-homogeneous), the process history at time t does not affect the instantaneous probability of events at time t. If covariates

are absent, time t is the only factor determining the intensity. In this case, the intensity function becomes,

$$\lambda(t) = \lim_{h \to 0} \frac{\Pr(\Delta N(t) = 1)}{h} = \rho(t).$$

The intensity function, $\rho(t)$ is also called the rate function.

The Poisson process has a number of useful properties that simplifies our models in subsequent sections:

- The process is a Markov process, i.e. the probability of an event in $(t, t+h)$ may depend on t but is independent of $H(t)$.
- For $t > 0$, $N(t)$ is a Poisson random variable with mean $\Lambda(t)$.
- If $(u_1, t_1]$ and $(u_2, t_2]$ are non-overlapping intervals, then $N(u_1, t_1)$ and $N(u_2, t_2)$ are independent random variables.
- If the process is homogeneous, then the inter-event times are independent exponential random variables with rate λ.

7.2.3 Recurrent event data model

Methods for the analysis of recurrent event data are usually specified through the intensity function. The commonly used intensity function are the counting process with the Cox-type rate function (Cox, 1972, 1975). An extension of the Cox model is the Andersen and Gill's counting process model (Andersen and Gill, 1982). In this model, the partial likelihood of the Cox model is extended such that each subject contributes the number of events they experienced over the study period to the likelihood. Like the Cox model, the assumption of proportional intensity is used to estimate model parameters. Under the proportional intensity assumption for two subjects with covariate values, x_1 and x_2, the ratio $\lambda_{x_1}(t)/\lambda_{x_2}(t)$ does not depend on time (is constant). A major draw-back of the Andersen-Gill model is that it assumes each event is independent (conditionally on covariates), i.e., it still retain the Poisson assumption of independence.

Extensions of the Andersen-Gill model without the Poisson-type assumption have been proposed in the literature (see Lin et al. (2000) and references therein). These models have the advantage of modeling correlation in the recurrent events within a subject. Although the models are semi-parametric and have some robustness to model mispecification, they are not suitable for use under the MAR missing data mechanism. Similarly, semi-parametric GEE based methods in longitudinal

data with missingness are not suitable under MAR assumption. The statistical analysis we adopt here is a parametric approach.

Poisson regression can be used to model homogeneous Poisson process. In what follows, we assume that events occur in continuous time and according to a Poisson process. For a Poisson random effect model, we consider an intensity function for a subject of the form

$$\lambda_x(t, \theta|U = u) = u\lambda_x(t, \theta), \tag{7.2}$$

where u is a realization of the gamma-distributed random variable U, with mean 1 and variance ϕ, λ_x is covariate dependent intensity function and θ are model parameters. This intensity function belongs to the conditional process, $N|U = u$, and not to the marginal process N. The extended model is then given by

$$\begin{aligned} N(t)|U \sim \mathscr{P}[\Lambda_x(t, \theta|U)] \quad \text{and} \quad U \sim \text{Gamma}(\phi^{-1}, \phi), \quad \text{and} \\ N(t) \sim \mathscr{NB}\Big(\frac{1}{\phi}, \frac{1}{1 + \phi\Lambda_x(t, \theta)}\Big), \end{aligned}$$

where $\mathscr{P}$ and $\mathscr{NB}$ stand for the Poisson and the Negative Binomial distributions respectively, and

$$\Lambda_x(t, \theta|U = u) = \int_0^t u\lambda_x(w, \theta)\mathrm{d}w = u\Lambda_x(t, \theta).$$

The marginal distribution of $N(t)$ is a negative binomial model and is obtained by integrating out the random effects from the mixed Poisson-Gamma distribution. The advantage of having a Poisson conditional process can be linked with the properties of Poisson process given in section 7.2.2. The process is memoryless and inter-event times can easily be simulated from the exponential distribution.

The contribution of a specific subject to the joint likelihood when, say, n

events occur at times $t_1, \ldots, t_n$, with U specified as above, is then given by

$$\begin{aligned}
L_N(\phi,\theta) &= \int f_{N(T),T_1,\ldots,T_n,U}(n,t_1,\ldots,t_n,u)\,\mathrm{d}u \\
&= \int f_{N(T),T_1,\ldots,T_n|U}(n,t_1,\ldots,t_n) f_U(u)\,\mathrm{d}u \\
&= \left[\prod_{j=1}^{n}\frac{\lambda_x(t_j,\theta)}{\Lambda_x(T,\theta)}\right] n! \int \frac{\exp\{-u\Lambda_x(T,\theta)\}(u\Lambda_x(T,\theta))^n}{n!} f_U(u)\mathrm{d}u \\
&= n!\left[\prod_{j=1}^{n}\frac{\lambda_x(t_j,\theta)}{\Lambda_x(T,\theta)}\right] \underbrace{\frac{\Gamma\left(n+\frac{1}{\phi}\right)}{n!\Gamma\left(\frac{1}{\phi}\right)}\left[\frac{\phi\Lambda_x(T,\theta)}{\phi\Lambda_x(T,\theta)+1}\right]^n\left[\frac{1}{\phi\Lambda_x(T,\theta)+1}\right]^{1/\phi}}_{f(\theta,\phi)},
\end{aligned} \tag{7.3}$$

where $f(\theta,\phi)$ denotes the probability mass function of a negative-binomial distributed random variable with mean $\Lambda_x(T,\theta)$ and variance $\Lambda_x(T,\theta)+\phi\Lambda_x^2(T,\theta)$.

In line with Cook and Lawless (2002), we assume that the treatments affect the intensity function through a multiplicative model of the form

$$\lambda_x(t,\theta) = u\lambda_0(t,\delta)g(x;\beta), \tag{7.4}$$

where $\lambda_0(t,\delta)$ is the baseline intensity function, $g(x;\beta)=\exp(\beta x)$ is a function of covariates and $\theta=(\beta,\delta,\phi)'$ is the parameter of interest. For convenience, we will use $g(x;\beta)=\exp(\beta x)$ since no restrictions are needed on the values of β, which can simply be interpreted as log-relative intensities. If we assume that the rate of events is constant over the study period, then $\lambda_0(.)$ is assumed specified up to a parameter $\delta>0$ (i.e. $\lambda_0(t,\delta)=\delta$). This yields a homogeneous Poisson process. A non-homogeneous Poisson process can be specified using, for example, a Weibull intensity function (i.e $\lambda_0(t,\delta)=\delta_0\,\delta_1\,t^{\delta_1-1}$), and can be monotone decreasing ($0<\delta_1<1$) or increasing ($\delta_1>1$). Other choices are discussed in Akacha and Benda (2010). If a constant rate is chosen, and we leave out the density of the negative-binomial in (7.3), we obtain $L_N(\phi,\theta)=T^{-n}$.

The likelihood function (7.3) can be used to model data sets with no missing cases (complete data), complete cases (with missing data but analyse only complete sequence) and completed cases (completed data through imputation).

Sometimes, a valid analysis for missing data can be accomplished by neither deleting nor imputing the missing data. In this case, all the available data are analysed in a likelihood framework (direct likelihood- DL), which is valid under an

ignorable missingness process (i.e. MAR with parameter separability (Carpenter et al., 2002)). A little modification of (7.3) yields the required likelihood, i.e.

$$L_{i,\text{obs}}(\phi,\theta) = n_{i,\text{obs}}!\left[\prod_{j=1}^{n_{i,\text{obs}}} \frac{\lambda_{x_i}(t_{ij},\theta)}{\Lambda_{x_i}(t_{id},\theta)}\right] \frac{\Gamma\left(n_{i,\text{obs}} + \frac{1}{\phi}\right)}{n_{i,\text{obs}}!\Gamma\left(\frac{1}{\phi}\right)} \left[\frac{\phi\Lambda_{x_i}(t_{id},\theta)}{\phi\Lambda_{x_i}(t_{id},\theta)+1}\right]^{n_{i,\text{obs}}} * \left[\frac{1}{\phi\Lambda_{x_i}(t_{id},\theta)+1}\right]^{1/\phi}, \tag{7.5}$$

where $n_{i,\text{obs}} = N_i(t_{id})$, and t_{id} is the dropout time for subject i. Our aim is to impute $n_{i,\text{mis}}$- the missing part of the recurrent event data sequence. Although DL is valid under an ignorable missing data mechanism, we adopt imputation to allow for flexibility (imputation models can be different from the analysis model) and transparency (missingness assumptions can be easily varied) in our sensitivity tool. Multiple imputation (MI) was introduced in the Bayesian framework (Rubin, 1987). Frequentist alternatives have also been proposed (Little and Rubin, 2002). We will consider two of these methods, imputation based on asymptotic normal properties of maximum likelihood estimators (MLE) and imputation based on a bootstrap approach. In large samples, the two are expected to be equivalent, although a bootstrap approach is to be preferred in small samples. These frequentist methods, like the Bayesian approach, are 'proper' imputation procedures because the ML estimates are asymptotically equivalent to a sample from the posterior distribution of the parameter. Basic characteristics that makes an imputation procedure to be proper can be found in Van Burren (2012). We introduce first the idea of imputing the missing recurrent data sequence based on the waiting time approach before discussing the imputation methods.

7.3 Methods of Imputation

Both the frequentist based imputation methods and the Bayesian method of MI will be introduced in section 7.3.2. The motivation for using the frequentist MI approach is that they are (approximate) proper imputation methods, simple to use and can easily be implemented by applied statisticians. The waiting times for imputation are generated using the unconditional counting process of the mixed effect Poisson process.

7.3.1 Waiting times or Gap times

In section 7.2.3, the data was assumed to be generated from a mixed effect Poisson process, with the random effects, $u_i's$, assumed to follow a gamma distribution. Conditional on the u_i, the event counts were assumed to follow a Poisson process with rate function $u\lambda_x(t,\theta)$. To impute the missing events, the unconditional counting process is required since we now work with the marginal counting process. The marginal intensity function is given as

$$\lambda_x(t,\theta|H_i(t)) = \left\{\frac{1+\phi N_i(t^-)}{1+\phi\Lambda_x(t,\theta)}\right\}\lambda_x(t,\theta), \tag{7.6}$$

where $N_i(t^-)$ is the number of events that occur in the interval $[0,t)$, and $H_i(t)$ is the history in that interval. The intensity function given by (7.6) at any time t depends both on ϕ and on the process history prior to t and is therefore no longer that of a Poisson process.

In order to use (7.6) to generate new waiting times, we consider the distribution of the waiting time, W_j between $(j-1)$st and jth events given by

$$\Pr\{W_j > w_j|T_{j-1}=t_{j-1},H(t_{j-1})\} = \exp\left\{-\int_{t_{j-1}}^{t_{j-1}+w_j}\lambda_x(u|H(u))\mathrm{d}u\right\}. \tag{7.7}$$

Equation (7.7) can be used to simulate a general intensity model. As an event occurred at t_{j-1}, then W_j for the jth event is simulated based on

$$B_j = \int_{t_{j-1}}^{t_{j-1}+w_j}\lambda_x(t|H(t))\mathrm{d}t, \tag{7.8}$$

where B_j has a standard exponential distribution (see Cook and Lawless (2007)). By repeating (7.8) for $j=1,2,\ldots$, successive event times $t_j = t_{j-1}+w_j$ can be

generated. For the constant rate, we have

$$\begin{aligned} B_j = & \int_{t_{j-1}}^{t_{j-1}+w_j} \left\{ \frac{1+\phi N_i(t^-)}{1+\phi\Lambda_x(t,\theta)} \right\} \lambda_x(t,\theta)\mathrm{d}t \\ = & \int_{t_{j-1}}^{t_{j-1}+w_j} \left\{ \frac{1+\phi N_i(t^-)}{1+\phi\Lambda_x(t,\theta)} \right\} \delta\exp(\beta x)\mathrm{d}t \\ = & \frac{1+\phi N(t^-)}{\phi} \int_{t_{j-1}}^{t_{j-1}+w_j} \frac{\phi\,\delta\exp(\beta x)}{1+\phi\,\delta\,t\exp(\beta x)}\mathrm{d}t \\ = & \frac{1+\phi N(t^-)}{\phi} \ln\Big(1+\phi\,\delta\,t_j\exp(\beta x)\Big)\Big|_{t_{j-1}}^{t_{j-1}+w_j}. \end{aligned} \tag{7.9}$$

If we make W_j the subject of the formula we get

$$W_j = \frac{\exp\left[\frac{B_j}{\frac{1}{\phi}+N(t_{j-1})} + \ln\Big\{1+\phi\,\delta\,t_{j-1}\exp(\beta x)\Big\}\right] - 1}{\phi\,\delta\exp(\beta x)} - t_{j-1}. \tag{7.10}$$

When $\phi = 0$, an equivalent of the waiting time given in (7.10) can be derived for a classical homogeneous Poisson process with intensity function $\lambda_x(t,\theta) = \delta\exp(\beta x)$. For this case, the waiting times are exponentially distributed with rate parameter $\delta\exp(\beta x)$. Using l'Hospital's rule ($l'HR$) for the calculation of the limits, we obtain

$$\begin{aligned} \lim_{\phi\to 0} W_j = & \lim_{\phi\to 0} \frac{\exp\left(\frac{B_j}{\frac{1}{\phi}+N(t_{j-1})} + \ln\Big(|1+\phi\,\delta\,t_{j-1}\exp(\beta x)|\Big)\right) - 1}{\phi\,\delta\exp(\beta x)} - t_{j-1} \\ \overset{l'HR}{=} & \lim_{\phi\to 0} \frac{\left(\frac{B_j}{\phi^2[\phi^{-1}+N(t_{j-1})]^2} + \frac{\delta\,t_{j-1}\exp(\beta x)}{1+\phi\,\delta\,t_{j-1}\exp(\beta x)}\right).\phi}{\delta\exp(\beta x)} - t_{j-1} \\ = & \frac{B_j + \delta\,t_{j-1}\exp(\beta x)}{\delta\exp(\beta x)} - t_{j-1} \\ = & \frac{B_j}{\delta\exp(\beta x)}, \end{aligned} \tag{7.11}$$

i.e. the gap times w_j, $(j = 1, 2, \ldots)$ between events are independent and identically distributed exponential random variables with rate $\delta\exp(\beta x)$.

In the motivating study, patients may drop out of the study because of drug related reasons, and patients are censored after the whole study period. To imple-

ment the proposed imputation method using the waiting time approach, a complete data set is generated by drawing random effects u from a gamma distribution and the rate function is computed using 7.4. The intensity function is used to draw the waiting times from the exponential distribution. If the waiting times are greater than the maximum time, the waiting time is censored. The number of events are then counted between the starting time and the censored time. MAR dropout is introduced in the complete data set by allowing dependence of the dropout model on the ratio of current number of events ($N(t_j)$) and time t_j (rate per unit time). That is, the probability, that a patient drops out after t_j is $p(t_j) = Pr(D_j = 1)$, where $D_j \sim Bernoulli[p(t_j)]$, and $logit[p(t_j)] = \beta_0 + \beta_1 x + \gamma \frac{N(t_j)}{t_j}$. The parameter, β_0 is varied in order to determine the percentage of missing data.

In order to impute the missing data, suppose patient i with count $N_i(t_{id})$ dropped out at time t_{id}. The waiting time approach implies $N(t_{id}) = N(t^-)$ and $t_{id} = t_{j-1}$. We follow the steps:

1. Let $j = 1$
2. The waiting time w_j between the event t_{id} and t_{id+j} is computed using (7.10)
3. Check if $t_{id+j} = t_{id} + \sum_{k=1}^{j} < T$
   (a) If $t_{id+j} \geq T$ stop
   (b) If $t_{id+j} < T$ then,
4. $N(t_{id+j}) = N(t_{id}) + j$
5. Let $j = j + 1$
6. Repeat 1-4
7. The final imputed time is t_{id+j} with corresponding count $N(t_{id}) + j$.

The DL function given in (7.5) can be used for the estimation of ϕ, δ and β needed in (7.10). This method results in single imputation of the missing data and the uncertainty in parameters used to impute are not taken into account. To avoid this, Rubin (1987) proposed multiple imputation (MI) as a flexible alternative to single imputation methods. MI solves the problem of uncertainty in the single imputation methods where imputation parameters are drawn from a conditional distribution. It is an extension of likelihood-based methods in that it adds an extra step in which imputed data values are drawn and final analyses combined. There has been a massive literature in support of MI (see Rubin (1996), Schafer (1997),

Collins et al. (2001) and Schafer and Graham (2002)). On the other hand, some authors have criticized MI on the grounds that Rubin's estimate of variance as too conservative (see for example, Wang and Robins (1998), Robins and Wang (2000) and Nielsen (2003)). This criticism has been shown not to invalidate MI procedures in general (Rubin, 2003).

Apart from Rubin's proper imputation there are various approximate methods of creating multiple imputations. These include, but are not limited to the use of posterior distribution from a subset of the data, refining approximate draws using importance sampling and drawing from pragmatic conditional distributions. Details of these methods can be found in Little and Rubin (2002). The imputation methods for this work focus on the use of the asymptotic distribution of the maximum likelihood (ML) estimates, and substitution of the ML estimates from bootstrapped samples. We introduce first Rubin's idea of MI.

7.3.2 Bayesian Multiple imputation

Rubin (1987) developed multiple imputation (MI) in the Bayesian framework. It is a simulation-based technique where missing values are replaced by $M > 1$ Bayesian draws from the conditional distribution of $N_{i,miss}$ given $N_{i,obs}$ and relevant covariates X_i, creating M completed data sets (*imputation phase*). Each of the M completed data sets are analysed using the same statistical procedures that would have been used had the data been complete (*analysis phase*). The M parameter estimates and their standard errors are then combined into a single set of results (*pooling phase*).

A convenient way to create multiple imputation in the imputation phase is to use data augmentation algorithm (Tanner and Wong, 1987; Schafer, 1997). This involves a two-step procedure that consists of an imputation step (I-step) and a posterior step (P-step).

- In the I-step, the missing data are drawn based on the observed data, covariates and the current parameter estimate
- The updated parameter estimate is drawn based on the current data in the P-step
- Under some conditions, the resulting imputed data sets define a Markov Chain which converges to the stationary distribution of $N_{i,mis}|N_{i,obs}, X_i$ for all $i \in \{1, \ldots, N\}$.

After the analysis of the M completed data sets, the results are pooled. The pooling process is as follows. Suppose the parameter estimates of θ from the M completed

data sets is $\tilde{\theta}_m$ and the variance V_m, then the MI estimator of θ is

$$\tilde{\theta}_{\mathrm{MI}} = \frac{1}{M}\sum_{m=1}^{M}\tilde{\theta}_m. \tag{7.12}$$

The estimate of the variance combines between and within imputation variability and is given by

$$\mathrm{V}_{\mathrm{MI}} = \frac{1}{M}\sum_{m=1}^{M}\mathrm{V}_m + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{m=1}^{M}(\tilde{\theta}_m - \tilde{\theta}_{\mathrm{MI}})^2. \tag{7.13}$$

Equations (7.12) and (7.13) are referred to as 'Rubin's rules for MI' and inference for θ is based on these equations. This gives,

$$\frac{\tilde{\theta}_{\mathrm{MI}} - \theta}{\sqrt{V_{MI}}},$$

which has an approximate t_ν distribution with

$$\nu = (M-1)\Big(1+\frac{W}{B}\Big)^2, \tag{7.14}$$

where $W = 1/M\sum_{m=1}^{M}\mathrm{V}_m$ and $B = 1/(M-1)\sum_{m=1}^{M}(\tilde{\theta}_m - \tilde{\theta}_{\mathrm{MI}})^2$. This apply for univariate parameter θ. Extensions to multi-dimensional estimators are straightforward (see Schafer (1997)). The so-called approximate proper imputations considered in this work are described below.

7.3.3 Asymptotic ML estimate

The asymptotic ML imputation methods use the asymptotic normal distribution properties of MLE's to draw imputation parameters from its asymptotic normal distribution. For recurrent event data, the imputation strategy for missing events follow the step-by-step approach laid out below:

1. Fit a negative-binomial model to the observed counts using direct-likelihood given in (7.5), extract $\hat{\theta} = (\hat{\phi}, \hat{\delta}, \hat{\beta})'$.

2. Consider the asymptotic distribution of $\theta \sim N(\hat{\theta}, \hat{\Sigma})$, where $\hat{\Sigma}$ is the inverse of the observed information matrix from the MLE of θ. Draw θ_k from this distribution and use it to draw new waiting times given by equation 7.10.

3. Complete the data set using the new waiting times.

4. Repeat 2 and 3 m times (where m can be greater than 100) to create m imputed data sets.

5. Fit the negative-binomial model to the completed data sets and save the point estimates and standard errors.

6. Combine results with Rubin's rules i.e. equations (7.12) and (7.13).

The advantage of this method is that it is very simple to use, and in large samples, it correctly propagates asymptotic uncertainty in the ML estimate of θ.

7.3.4 Bootstrap imputation method

The bootstrap approach is an alternative to the asymptotic ML method. Random samples, with replacement, of the same size n as the observed sample are taken and the DL method is used to estimate parameters. The ML estimates from the bootstrap samples are asymptotically equivalent to samples from the posterior distributions of the parameters and do not rely heavily on large-sample properties. The algorithm for imputation with recurrent event data is similar to the one laid out in section 7.3.3, but with steps 1 and 2 replaced by non-parametric bootstrap procedures, i.e.

1. Draw a bootstrap sample with replacement of size n from the original data

2. Fit a negative-binomial model to each sample and combine the estimates to form $\tilde{\theta} = (\tilde{\phi}, \tilde{\delta}, \tilde{\beta})'$. Use $\tilde{\theta}$ to draw new waiting times and complete the data using the new waiting times.

The imputation steps described above use the assumption that patients that share the same observed data and covariates (e.g treatment) will have the same statistical behavior in their future evolution whether they are observed or not. This is an MAR assumption, and it implies that patients who drop out under the active treatment arm, will be imputed using information from this group. In many clinical trial settings however, the MAR assumption may be unrealistic. For instance, patients may discontinue treatment due to adverse treatment effects, lack of efficacy or some MCAR missing data mechanism related reasons. To analyze this data, a realistic imputation assumption is to impute missing data for the treated group using information from the placebo arm. For patients with missing data in the placebo arm, they are imputed using the event rate in the placebo arm. This is because the MAR assumption is realistic for the placebo arm since patients in the arm received no active substance in the first instance. We investigate this and other plausible scenarios in a simulation study.

7.4 Simulation

We assess the performance of the asymptotic and bootstrap methods of imputation using the waiting times approach in a simulation study. In addition, since the imputation method was developed under the mixed gamma-Poisson model, we investigate the impact of using different data generation process in recurrent event data settings. The data for the motivating example is from a placebo-control trial. This motivated the investigation of the impact of imputation of missing data in the treated arm using the information from the placebo arm.

Sometimes however, clinicians might know, based on experience, a realistic percentage increase in event rate for patients who discontinued treatments. We therefore imputed data for patients who discontinued in the active arm with higher event rates than the MAR rate of $\lambda_{\text{trt}}(t)$ i.e. from

$$\lambda_{\text{new,trt}}(t) = \lambda_{\text{trt}}(t) * k, \quad k \in \{1.05, 1.10, 1.20, 1.50\}. \tag{7.15}$$

In clinical settings, the constant rate rather than the Weibull is usually assumed for the intensity function. We therefore assume the former for this study. As mentioned earlier, two treatment groups are compared: the active arm and the placebo arm. We set a constant treatment effect of $\beta_1 = -0.3$, with maximum follow up time $T = 112$ days. The random effect variance, ϕ, is fixed at 0.5, and the decay rate, δ at 0.02 for the mixed Poisson process. The mean rate of events per unit time in the treated arm is 0.015 (i.e. $0.02 * \exp(-0.3)$) while that of the placebo arm is 0.02 since we expect treatment intervention to reduce event rates in the former. The mean over the study period is 1.66.

The first simulation settings involves the Comparison of Asymptotic and Bootstrap imputation methods using 10, 20, 50 and 100 imputations for the former 10, 20 and 50 imputation for the latter in small ($n = 100, 200$) and large ($n = 400, 1000$) sample sizes. Missingness percentage considered is 30% treated vs. 30% placebo. The further simulations itemized below used 10 imputations for both methods, and larger sample size as this was clearly sufficient.

- Evaluation of impact of missingness percentage in the Asymptotic and Bootstrap methods: 30% missingness in treated arm vs. 10% missingness in placebo arm and 40% missingness in treated arm vs 10% missingness in placebo arm.

- Evaluation of the impacts of using different random effects for data generation. Uniform, (U[-1,1.5]) and Normal, ($N[0, 1/2]$) random effects are used. These choice ensure that the variance of the random effects is close to 0.5, which

is the choice used for the Gamma-Poisson mixture. Missingness percentages considered are 30% missingness in treated arm vs. 30% missingness in placebo arm.

- Evaluation of other event generation process. The processes considered are Poisson, Weibull, Conditional, and Autoregressive process see (Metcalfe and Thompson, 2006; Jahn-Eimermacher, 2008). Missingness percentage considered is 30% missingness in treated arm vs. 30% missingness in placebo arm.

- Imputation of missing data in the treated arm using event rate of the placebo arm. The data generation follows the gamma-Poisson mixture model described in section 7.2.3. Missingness percentage considered are 30% missingness in treated arm vs. 30% missingness in placebo arm, 30% missingness in treated arm vs. 10% missingness in placebo arm and 40% missingness in treated arm vs 10% missingness in placebo arm.

- Imputation of missing data in the treated arm using rates higher than the MAR rates as given in (7.15). The gamma-Poisson mixture model is used for the data generation.

7.4.1 Asymptotic and Bootstrap simulation

A unique feature of MI is that it provides a mechanism for dealing with inherent uncertainty of imputations themselves. The question of how many imputations are needed has been discussed in the literature. Graham et al. (2007) approached the problem in terms of loss of power for hypothesis testing. Using simulation study, they recommend 20 imputations for 10%-30% missing information (the percentage of missing information is $W/(W+B)100\%$, where W and B are as defined in section 7.3.2), and 40 imputations for 50% missing information. For the current study, we assess the impact of the number of imputations Asymp $\in \{10, 20, 50, 100\}$ under the asymptotic method and Bootstrap $\in \{10, 20, 50\}$ under the bootstrap method. Bias and MSE in small and large sample sizes and 30% missing data in the two treatment arms is used to quantify their relative performance.

Table 7.3 shows the results when using 'many' asymptotic and bootstrap imputations. The data is generated from the gamma-Poisson process (see section 7.2.3). The row labeled NM is for the original complete data set (before introducing missingness) and DL gives the results of the direct likelihood approach. There is no gain in having sample size of 200 over 100 in terms of the bias for the treatment effect β. However, ϕ and δ have less bias with sample size of 200. Although absolute

bias is larger for $n = 200$, the relative bias (to NM) is smaller. Similar observation can be seen when large samples are used, i.e. the bias in β for sample size 400 is consistently lower than that of 1000. Overall however, the bias in β is smaller when large sample sizes are used than when small sample sizes (100 and 200) are used. Of course as expected, the MSE is consistently smaller as the sample size increases regardless of whether the sample is small or large.

Similarly, the Bias and the MSE suggest that the use of 10 imputations is sufficient under the asymptotic mle imputation and the bootstrap approach regardless of the sample size.

7.4.2 Effects of fraction of missing information on treatment estimates

In some settings, it is possible to have higher rate of drop outs in the treated arm than the placebo arm due to adverse effects. We consider settings where 30-40% of the observations are missing in the treated arm and about 10% missing data in the placebo arm. Table 7.4 presents the results for the setting. As expected, there is slight bias in treated estimates with increased amount of missing information in the treated arm. Under MAR assumption with sample size and number of imputation approaching infinity, we would expect to see no bias.

7.4.3 Event generation based upon alternative random-effects distributions

Gamma distribution is often used as random effects for mixed Poisson process because it leads to a closed form expression in the marginal process. Mixture distributions other than gamma may be assumed for the random effects. In this study, we consider realizations U from a uniform $U[-1, 1.5]$ distribution (with associated rate $\delta \exp(U + \beta x)$), and realizations Z from a normal distribution with mean zero and variance $1/2$ (with rate $\delta \exp(Z + \beta x)$). The choices ensure that the variance of the random effects is close to 0.5 which was chosen for the gamma-Poisson mixture. Table 7.5 shows the results of using alternative random effects distributions in the data generation process. The performance of the two imputation methods are similar and the bias in sample size of 1000 is negligible. Comparing this result with equivalent part of Table 7.3 suggests that the choice of the random effects that generated the data is immaterial even when the negative binomial model is fitted to the completed data sets.

Table 7.3: Bias and MSE in estimated treatment effect with 30% missing data in both placebo and treated arm: Asymptotic and Bootstrap imputations. Simulation results (multiplied by 10,000).

		Bias			MSE		
		ϕ	δ	β	ϕ	δ	β
	NM	-198	0	5	279	0.0	408
n=100	DL	-92	2	-36	511	0.0	603
	Asymp_{10}	33	6	-60	461	0.0	754
	Asymp_{20}	53	6	-66	446	0.0	761
	Asymp_{50}	75	6	-59	434	0.0	767
	Asymp_{100}	86	6	-58	432	0.0	762
	Bootstrap_{10}	-583	-1	-41	572	0.0	779
	Bootstrap_{20}	-583	-1	-37	583	0.0	779
	Bootstrap_{50}	-581	-1	-34	586	0.0	778
	NM	-123	-0	60	128	0.0	213
n=200	DL	-27	1	57	247	0.0	311
	Asymp_{10}	29	3	92	252	0.0	424
	Asymp_{20}	28	3	100	245	0.0	424
	Asymp_{50}	36	3	105	246	0.0	433
	Asymp_{100}	39	3	102	245	0.0	430
	Bootstrap_{10}	-203	0	89	264	0.0	445
	Bootstrap_{20}	-202	0	89	264	0.0	445
	Bootstrap_{50}	-204	0	88	260	0.0	445
	NM	-56	1	15	64	0.0	97
n=400	DL	-45	1	12	124	0.0	137
	Asymp_{10}	0	2	7	125	0.0	204
	Asymp_{20}	0	2	9	121	0.0	203
	Asymp_{50}	-1	2	8	120	0.0	201
	Asymp_{100}	1	2	7	120	0.0	203
	Bootstrap_{10}	-101	1	0	126	0.0	209
	Bootstrap_{20}	-116	0	4	129	0.0	211
	Bootstrap_{50}	-116	0	5	128	0.0	210
	NM	-60	0	-28	28	0.0	42
n=1000	DL	-37	1	-25	50	0.0	58
	Asymp_{10}	-7	1	-30	49	0.0	86
	Asymp_{20}	-5	1	-28	49	0.0	86
	Asymp_{50}	-7	1	-27	48	0.0	86
	Asymp_{100}	-6	1	-28	48	0.0	86
	Bootstrap_{10}	-44	1	-27	50	0.0	91
	Bootstrap_{20}	-43	1	-18	51	0.0	90
	Bootstrap_{50}	-43	1	-16	49	0.0	92

Table 7.4: Bias and MSE in estimated treatment effect with 30% and 40% missingness in the treated arm. Percentage of missing data in placebo arm is fixed at 10%. Simulation results (multiplied by 10,000).

		Bias			MSE		
Size(Missing)		ϕ	δ	β	ϕ	δ	β
	NM	-56	1	15	64	0.0	97
n=400 (30%)	DL	-51	1	9	97	0.0	127
	Asymptotic	-31	1	24	92	0.0	144
	Bootstrap	-94	0	9	95	0.0	146
	NM	-60	0	-28	28	0.0	42
n=1000 (30%)	DL	-54	0	-10	39	0.0	52
	Asymptotic	-35	0	7	37	0.0	58
	Bootstrap	-60	0	-1	38	0.0	59
	NM	-56	1	15	64	0.0	97
n=400 (40%)	DL	-40	1	24	105	0.0	134
	Asymptotic	-17	1	50	100	0.0	159
	Bootstrap	-93	0	16	102	0.0	162
	NM	-60	0	-28	28	0.0	42
n=1000 (40%)	DL	-56	0	-8	42	0.0	60
	Asymptotic	-32	0	22	40	0.0	69
	Bootstrap	-58	0	14	13	0.0	69

Table 7.5: Bias and MSE in estimated treatment effect with 30% missingness in both placebo and treated arm: Uniform and Normal random effects. Simulation results (multiplied by 10,000).

			Bias			MSE		
R.E	Size		ϕ	δ	β	ϕ	δ	β
U[-1,1.5]	n=400	NM	14	129	11	38	2	81
		DL	254	133	-6	68	2	115
		Asymptotic	272	134	25	68	2	137
		Bootstrap	233	133	28	68	2	145
	n=1000	NM	13	129	12	15	2	34
		DL	209	132	10	28	2	47
		Asymptotic	230	133	6	28	2	57
		Bootstrap	213	132	9	27	2	58
N[0,1/2]	n=400	NM	357	57	-11	75	0.0	107
		DL	-248	52	-39	92	0.0	130
		Asymptotic	-220	52	-29	88	0.0	183
		Bootstrap	-283	51	-31	95	0.0	190
	n=1000	NM	357	56	8	38	0.0	42
		DL	-275	51	14	43	0.0	54
		Asymptotic	-252	51	13	40	0.0	80
		Bootstrap	-278	51	9	43	0.0	82

7.4.4 Alternative event generation process

In practice, the actual process that generated the data is unknown. We consider four processes that may generate recurrent event data and created MAR missingness in the data. The imputation method under the mixed Poisson process described in section 7.3.1 is used to complete the data.

Poisson process

In a very unlikely situation where it is thought that patient specific heterogeneity is unnecessary in the model, the data can be obtained from a Poisson model. Events in Poisson process occur independently of one another, both between subjects and within each subject. The data generation is similar to the one described in section 7.2.3 for the mixed Poisson process but with the omission of random effect u. Waiting times are simulated as independent realization of an exponential distribution with rate $\lambda = \delta \exp \beta x$.

Weibull model

There are situations where an individual is particularly susceptible to further events shortly after a previous event. In this case the intensity function for further event will change over time and a Weibull model can be used to describe the waiting times between events. Its intensity function can be written as $u\,\delta_0\,\delta_1\,t^{\delta_1-1}\exp\beta x$, where δ_0 & δ_1 are positive. The intensity function is monotone decreasing when $0 < \delta_1 < 1$, constant when $\delta_1 = 1$ amd monotone increasing when $\delta_1 > 1$. Unlike the mixed Poisson process whose inter-arrival times are exponentially distributed, the waiting time for the Weibull intensity function has to be calculated directly using inverse CDF method. This is given by

$$w = \left(\frac{\ln(1-\nu)}{-U\delta \exp(\beta x)} + t^{\delta 1} \right)^{1/\delta_1} - t,$$

where $\nu \sim \text{Uniform}(0,1)$, t is previous event time and δ_1 is the shape parameter that describes how the intensity of an event is distributed across time. For our simulation, we take $\delta_1 = 0.9$ and $\delta = 0.032$ so that the expected number of events is similar to that of the mixed Poisson process. Our choice of δ_1 reflects a declining intensity function.

Autoregressive model

In settings where the occurrence of an event in a patient makes it more likely for the patient to have further events, an underlying autoregressive process can be assumed. For instance, in an epileptic study, occurrence of a seizure in a patient can make it more likely for the patient to experience further seizures. For this study, we generated the waiting times using

$$w = \frac{-\ln \nu}{\exp\{-0.1 + \beta x + 0.1n\}},$$

where n is the number of previous events. The quantity $exp\{-0.1 + \beta x + 0.1n\}$ is the rate parameter and was chosen such that the number of events observed over the study period is similar to other data generation process.

Conditional process

The conditional process is similar to the autoregressive model because it makes the assumption that event rates may increase or decrease after observing a certain number of events in a patient. Unlike the autoregressive model, the number of events does not contribute to increase or decrease in the rate of events in a continuous manner. For this study, we consider an extended mixed Poisson process such that the rate parameters

$$\lambda_1 = u\,\delta \exp(\beta x) \text{ and } \lambda_2 = u\,\delta \exp(\beta x + 0.1)$$

are used when event counts is less than 2 and greater than or equal to 2 respectively. That is, a patient will have a slightly increased chance of having subsequent events after experiencing the first event. Both u and δ are as defined for the mixed Poisson process.

Table 7.6 gives the proportion of events observed in the treatment group when subjects are censored at 112 days. The simulation is based on 1000 replications for each model. The mixed Poisson process, the Weibull model and the conditional model result in greater proportion of subjects with no events or more than 4 events compared to other process. This reflects the larger variance of models especially when compared with the Poisson model (the mixed-Poisson, the Weibull and the conditional models incorporated random effects which introduce extra variability than the Poisson model in order to capture over-dispersion). The Weibull model was constructed to have expected number of events over the study period to be equivalent to the mixed-Poisson process, hence the similarity in the result obtained

Table 7.6: Proportion of events observed in treatment group using simulated data for the models, n=1000, 1000 replications and censoring at 112 days.

Num. of events	0	1	2	3	4	5	6	7	8	> 9
Poisson	0.190	0.315	0.262	0.146	0.060	0.020	0.005	0.001	0.000	0.000
Mixed Poisson	0.298	0.271	0.185	0.112	0.063	0.035	0.018	0.009	0.005	0.005
Weibull	0.298	0.272	0.185	0.111	0.063	0.034	0.018	0.009	0.005	0.005
Autoregressive	0.222	0.309	0.237	0.132	0.061	0.025	0.009	0.003	0.001	0.000
Conditional	0.299	0.271	0.173	0.110	0.066	0.037	0.021	0.011	0.006	0.007

in the table. Table 7.7 is the result of applying our imputation methods to the data generated through the Weibull, the conditional, the Poisson and the autoregressive model. The bias in the treatment effects is small and the result is comparable with the equivalent results from Tables 7.3. However, there is a little bias in the results based on the data generated from the conditional model.

Note that the entry for ϕ is not included for the Poisson and the autoregressive models. This is because the two models did not include random effects. There were numerous numerical instability in an attempt to force imputation strategy based on mixed effect Poisson process on the two data generation process because $\phi \simeq 0$. When the sample size is 400, about 60% of the simulated data experienced numerical problems under the Poisson data generation process with the asymptotic imputation and almost 35% failed under the bootstrap procedure. There are fewer cases of non-convergence with sample size of 1000. The performance of the imputation method is better under the autoregressive models as only 39 samples out of 1000 replications failed to converge under the autoregressive model with sample size of 1000. The percentage with non-convergence increases with the number of imputation. Although in reality, there will almost always be subject specific effect in a recurrent event data set, the presence of numerical instability, as we observe with the Poisson process, could be a pointer to the fact that the data is better imputed and analyzed in the Poisson process framework than the mixed Poisson process settings.

A method to solve the numerical instability when imputing the data under the mixed-Poisson process is to use 'conditional imputation'. By this we mean, a threshold is set for ϕ (say for instance, $\phi = 0.001$). If ϕ is less than this threshold, the waiting time under the Poisson model (see equation (7.11)) is used to impute the missing data. Otherwise, the waiting time of the mixed Poisson process is used.

Table 7.7: Bias and MSE in estimated treatment effect under the Weibull, Conditional, Poisson and Autoregressive data generation process. Imputation was done under mixed Poisson process. Simulation results (multiplied by 10,000).

			Bias			MSE		
Model	Size		ϕ	δ	β	ϕ	δ	β
Weibull	n=400	NM	-46	-120	-19	62	1	103
		DL	594	-112	-21	181	1	147
		Asymptotic	521	-112	-66	182	1	188
		Bootstrap	418	-113	-57	170	1	192
	n=1000	NM	-29	-120	-34	28	1	40
		DL	538	-113	-36	87	1	58
		Asymptotic	520	-113	-11	88	1	66
		Bootstrap	470	-114	-14	83	1	68
Conditional	n=400	NM	561	9	-104	102	0.0	109
		DL	432	8	-104	132	0.0	140
		Asymptotic	450	8	-118	148	0.0	163
		Bootstrap	366	8	-115	140	0.0	166
	n=1000	NM	572	9	-129	61	0.0	49
		DL	458	8	-135	67	0.0	63
		Asymptotic	477	8	-128	76	0.0	74
		Bootstrap	435	8	-131	73	0.0	75
Poisson	n=400	NM	-	0	1	-	0.0	49
		DL	-	3	-18	-	0.0	80
		Asymptotic	-	59	21	-	13	164
		Bootstrap	-	3	33	-	0.0	133
	n=1000	NM	-	0	5	-	0.0	19
		DL	-	1	4	-	0.0	28
		Asymptotic	-	29	-9	-	1	73
		Bootstrap	-	1	-12	-	0.0	52
Autoregressive	n=400	NM	-	4	-292	-	0.0	75
		DL	-	0	-234	-	0.0	94
		Asymptotic	-	9	-206	-	1	411
		Bootstrap	-	-1	-249	-	0.0	108
	n=1000	NM	-	4	-343	-	0.0	36
		DL	-	0	-282	-	0.0	40
		Asymptotic	-	2	-289	-	0.0	47
		Bootstrap	-	-1	-293	-	0.0	21

7.4.5 Imputation under MNAR assumption- Treated follows Placebo

Table 7.8 contains the results of imputing the missing data in the active arm using the event rates of placebo arm. If the MAR assumption is used, the estimated treatment effects will be close to -0.3 when in actual fact the effect is higher than -0.3 as demonstrated by the imputation. As the fraction of missing data increases in the treated arm, the rate of events also increases, which is intuitively reasonable since a larger percentage of patients in the treated arm now follows the higher placebo arm rate.

7.4.6 Imputation under MNAR assumption- Higher event rates than MAR assumption

Tables A.3-A.5 in section A.5 of Appendix A show the results of incrementing the event rates in the active arm by 5%, 10%, 20% and 50% for imputing the missing recurrent events in the active arm with varying percentage of missing data. The impact of 5-10 percent increase in the event rates does not appear to increase the estimate of the treatment effect by larger percentage. However, the impact of using higher rate for imputation became pronounced when the imputation was done with 50% increment in event rate than the MAR assumption. The degree of increment in practical settings is a matter of judgement from clinical experts.

7.5 Application to Bladder Cancer Trial

In the paper of Wei et al. (1989), marginal approach to the analysis of multivariate failure time based upon an elaboration of the Cox proportional hazards model was proposed. In the current work, the average treatment effect captured by a parameter is considered. Since the time to the end of the study was not explicitly stated, we base our work on the assumption that the end of the study is equivalent to the maximum follow-up time, which is 64 months. Apart from the treatment effects β, we also adjusted for the two baseline variables in the model. All the models fitted to this data are based on the assumption of homogeneous mixed Poisson process and we define ϕ as the random effect variance and δ as the decay rate. We adopt a mixed-Poisson process approach because the variance (2.315) of the count is much larger than its mean (1.318). In addition, 20 imputations are used for the asymptotic and the bootstrap methods and 500 bootstrap samples from the original data is used for each imputation under the bootstrap method. We used 20 imputations even though 10 imputations were adjudged adequate in the simulation study because the sample

Table 7.8: Imputation of treated arm using placebo rate $\lambda_p(t)$. A and P stand for active and placebo arms respectively.

			Parameters			Std. Err.		
Missing	Size		ϕ	δ	β	ϕ	δ	β
		True Val.	0.5000	0.0200	-0.3000	-	-	-
30% A-30% P	n=400	Asymptotic	0.5618	0.0202	-0.2129	0.1701	0.0710	0.1075
		Bootstrap	0.5519	0.0201	-0.2136	0.1545	0.0707	0.1031
	n=1000	Asymptotic	0.5632	0.0201	-0.2151	0.1047	0.0449	0.0676
		Bootstrap	0.5587	0.0201	-0.2151	0.0952	0.0448	0.0654
30% A-10% P	n=400	Asymptotic	0.5600	0.0201	-0.2109	0.1585	0.0709	0.1074
		Bootstrap	0.5528	0.0200	-0.2133	0.1529	0.0707	0.1029
	n=1000	Asymptotic	0.5612	0.0200	-0.2109	0.0986	0.0449	0.0677
		Bootstrap	0.5585	0.0200	-0.2118	0.0949	0.0448	0.0652
40% A-10% P	n=400	Asymptotic	0.5704	0.0201	-0.1893	0.1581	0.0713	0.1092
		Bootstrap	0.5641	0.0200	-0.1909	0.1511	0.0711	0.1032
	n=1000	Asymptotic	0.5704	0.0200	-0.1894	0.0986	0.0451	0.0689
		Bootstrap	0.5673	0.0200	-0.1902	0.0940	0.0450	0.0653

size for the data is smaller than the ones considered in the simulation study.

Table 7.9 is the result of fitting the DL model and the use asymptotic and bootstrap imputation methods to complete the bladder cancer data under the MAR assumption. The standard errors under the bootstrap method of imputation are consistently smaller than the DL and the asymptotic methods. This is expected if we consider the small size of the data set. There is a significant difference between the two treatment arms (β) at 5% level of significance under the bootstrap approach. However, a non-significant treatment effects are obtained under the DL and asymptotic imputation method. A possible explanation for this is that the precision of the estimates is reduced when draws are taken using the asymptotic MLE properties because of the small sample size. Of course small sample size also affects standard errors in DL approach. This is not the case with the bootstrap method and as such, we are inclined to conclude that inference based on this estimate is more likely to be representative of the unknown true value.

We also explore the idea of imputing the missing data in the treated arm using the placebo event rate. A possible reason for patients to discontinue treatment in the bladder cancer study is death. A patient that dies received no treatment afterwards, and it is logical to assume that their event rate will be similar to the placebo arm event rate after death. Table 7.10 is the result of imputing the data using the

Table 7.9: Fit of Direct Likelihood, Asymptotic Imputation and Bootstrap Imputation to the bladder cancer data.

	DL			Asymptotic			Bootstrap		
	Estimate	S.E.	p-value	Estimate	S.E	p-value	Estimate	S.E	p-value
ϕ	0.4981	0.2544	0.0537	0.5982	0.2719	0.0335	0.4487	0.1911	0.0217
δ	0.0353	0.0111	0.0020	0.0310	0.0099	0.0031	0.0291	0.0074	0.0002
β	-0.4605	0.2707	0.0928	-0.4263	0.2781	0.1330	-0.4727	0.2183	0.0338
num.	0.1877	0.0735	0.0125	0.1926	0.0799	0.0206	0.2131	0.0592	0.0006
size	-0.0337	0.0921	0.7151	-0.0120	0.0977	0.9032	-0.0128	0.0744	0.8644

placebo arm event rate. In this case, the treatment effect becomes non-significant under the two models. This is intuitive since there are high degree of missingness in the treated arm.

Table 7.11 is the results of imputing the missing data in the treated arm using event rate higher than the MAR. We consider 5, 10, 20 & 50% increase in the MAR rate as was done in the simulation study but only focus on imputation based on bootstrap (since its performance is generally more reliable in small samples). When the rate is 20% and above, the treatment effect becomes non-significant. As we remarked earlier, the degree of increase in the rate than MAR is a matter of clinical judgement.

Table 7.10: Fit of Asymptotic Imputation and Bootstrap Imputation to the bladder cancer data using event rates in the placebo arm to impute data in the treated arm.

	Asymptotic			Bootstrap		
	Estimate	S.E.	p-value	Estimate	S.E	p-value
ϕ	0.6185	0.2605	0.0218	0.4069	0.2155	0.0636
δ	0.0313	0.0100	0.0030	0.0316	0.0076	0.0001
β	-0.2692	0.2483	0.2839	-0.2994	0.2126	0.1641
num.	0.1838	0.0715	0.0135	0.1874	0.0543	0.0010
size	-0.0067	0.0984	0.9457	-0.0295	0.0728	0.6867

7.6 Summary

The first simulation scenario compares the use of asymptotic and bootstrap methods for multiple imputation and the effects of the number of imputations used on their performances. The parameter of interest is the treatment effect (β), although other parameters in the model are also reported. Our results suggested that the use of 10

Table 7.11: Fit of Bootstrap Imputation to the bladder cancer data using higher rate than the MAR rate. Bold face entries are significant at 5% level of significance.

	$\lambda = 1.05$		$\lambda = 1.10$		$\lambda = 1.20$		$\lambda = 1.50$	
	Estimate	S.E.	Estimate	S.E	Estimate	S.E	Estimate	S.E
ϕ	**0.4443**	**0.1924**	**0.4332**	**0.1819**	**0.3692**	**0.1780**	**0.3494**	**0.1464**
δ	**0.0300**	**0.0080**	**0.0307**	**0.0081**	**0.0305**	**0.0072**	**0.0311**	**0.0075**
β	**−0.4687**	**0.2164**	**−0.4324**	**0.2157**	-0.3621	0.2072	-0.2115	0.2031
num.	**0.2089**	**0.0603**	**0.2022**	**0.0594**	**0.1881**	**0.0590**	**0.2158**	**0.0504**
size	-0.0213	0.0777	-0.0263	0.0800	-0.0130	0.0756	-0.0477	0.0765

imputations are sufficient under the methods when up to 30% values are missing in the two treatment groups. A pilot simulation with roughly 50% missing counts in the groups also affirmed this. We found that sample size (N) of 400 patients (200 in each arm) are sufficient to achieve this conclusion.

We assumed that the underlying process is a gamma- mixture of Poisson processes, with negative binomial regression model as standard analysis method. It is almost impossible in practice to know the actual process that generated the observed data. These included the nature of random effects and the actual event generation processes. Our simulation results found that the use of asymptotic and bootstrap imputation methods do not affect estimation of β, although its precision improved with $N = 1000$. It turns out that the imputation techniques are not affected by making a gamma random effects assumption when in fact the true random effects are uniformly or normally distributed. This is in line with simulation results in the literature that analyzed complete data (see Metcalfe and Thompson (2006)). When events are generated from Poisson and autoregressive models, we experienced numerical instability because of the attempt to force our imputation methods based on gamma-Poisson process on data generation methods with no random effect, i.e, $\phi = 0$. In practice, this numerical problem can signal that the gamma-Poisson model is not appropriate for the data at hand, and alternative models should be investigated.

The key assumption under the pMI technique is that patients in the active arm do not benefit from treatment after discontinuation. This is responsible for the increase in estimated event rates in the treated arm after pMI. The event rate increases further as the fraction of missing data increases. The choice of realistic higher event rates than the MAR to use for imputation in treated arm depends on the nature of the study, and it is determined by clinical experts, while avoiding data snooping.

Application to the bladder cancer data showed that parameter estimates under the direct-likelihood, asymptotic and bootstrap imputation methods are similar, as expected, under MAR missingness process. Due to small sample size, we used the small-sample degrees of freedom proposed by Barnard and Rubin (1999) for the p-value calculations in Tables 7.9- 7.11 under the frenquetist imputation methods. This method ensures that the degrees of freedom under multiple imputation does not exceed the complete data degrees of freedom (85-5=80 in the bladder cancer data).

Chapter 8

General Conclusions and Future Research

8.1 Conclusion

This thesis is concerned with methods of dealing with missing data. The first part unifies missing data problem into a distributional framework, while methods for imputation of missing data in recurrent event data is proposed and evaluated in a simulation study in the second part.

In chapter 2, we reviewed relevant literature on skew distributions and their unified class, the closed skew-normal (CSN) distribution. The link between sample selection and skew distributions were established. This link formed the central theme of the ideas discussed in chapters 3-6. The MINT trial data, which was used as motivating examples in the first part of the thesis, was explored. Results from the exploration showed that the data is skewed marginally. The skewness can be due to several factors, including but not limited to, the boundedness of the scores or non-ignorable missingness (sample selection). We also looked briefly at concepts of missing data, which cut across the two parts of the thesis.

In chapter 3, we used complete case analysis for the subjects that completed the trial. The missing data problem was treated as a hidden truncation problem, and as such, the use of skew-normal models are justified. The SN and ESN models are standard and well known in the statistical literature. We introduced a three-parameter skew distribution which we referred to as EGSN. This model has two parameters that control skewness and a third parameter which is a shift parameter. The SN, ESN and EGSN models were used in a simulation study where the data sets were generated in a sample selection settings but with bivariate skew-normal errors.

There is bias in parameter estimates as expected from complete case analysis, and the bias is more pronounced when the skewness parameter, $\lambda = 0.5$. The three models resulted in similar fits to the NDI scores, although the SN model may be preferred because it has fewer parameters. The bounds in the scores were adjusted for using truncated skew-normal distribution. We concluded chapter 3 by noting that complete case analyses failed to adjust for covariate information that led to non-response. The implication of this is inflated type-1 error, and this motivated the model developed in chapter 4.

In chapter 4, we developed a sample selection model with an underlying bivariate skew-normal distribution. This model has the advantage of adjusting for additional information about the non-response process, and we showed how this circumvented non-identifiability of the ESN model in chapter 3. Finite sample properties of the MLEs of the model were studied in a simulation study. The performance of our model was superior to classical Heckman's models. The bias observed in parameter estimates when $\lambda = 0.5$ in the models of chapter 3 was also observed in this case. This is due to the models inability to distinguish between $\lambda = 0$ and $\lambda = 0.5$, as there is stationarity of profile likelihood of λ at $\lambda = 0$. The treatment effect is not significant in all the models in chapter 4, even though it was significant in the models in chapter 3. This further buttresses the danger of complete case analysis under non-ignorable non-response.

The model in chapter 5 is a multilevel extension of the model discussed in chapter 4. Although, the developments of multilevel sample selection models are not new in the literature, the work we presented here is probably the first time it was linked to the CSN distribution. This link provided a framework that simplified the derivation of conditional mean and variance of the model, and was used to generalize Heckman two-step method to multilevel selection process. We focused on likelihood estimation of the parameters in the model which is rarely used in the literature. This is due in part to sensitivity of the approach to normality assumption. Unit and item non-response in the NDI scores were jointly analyzed. We noted the large standard error and high p-value for the sex variable in the item level equation in Table 5.4 relative to the same statistics for other variables. This could be an indication of numerical issue of the optimization routine. We will assess in our future work how robust our results are to changes in model specification and develop tools for sensitivity analysis for multilevel selection models.

The models presented in chapter 6 were based on the use of Gaussian copulas in sample selection settings with skew-normal marginals for the outcome equation. The principal contribution in this chapter is the use of the link between sample

selection and skew distributions to derive copula-based sample selection models. This gave a straightforward approach for the use of any differentiable copulas in this setting. The flexibility of copulas allowed us to model bounded outcomes using truncated skew-normal distributions as marginals. Our reference distribution for the outcomes is the asymmetric subfamily of the sinh-arcsinh distribution, which we referred to as SHASH model. Tractability and inferential advantages were the motivation for the use of the SHASH model over the Azzalini's SN model. Application to the NDI scores showed that the adjustment for the bounds in the outcome using truncated SHASH and SN models resulted in nonsignificant skewness parameter, even though the skewness parameters in the non-truncated counterparts were significant. A Gaussian copula was used on the ground that the association parameter is easily conceivable by applied researchers, and to establish links between results in chapter 6 and earlier chapters. We therefore caution on over-interpreting these results, knowing fully well that copulas can be abused, especially the Gaussian copula.

The second part of this thesis is motivated by a placebo-controlled trial which explores recurrent event data over a period of several weeks. We focus on a situation where the number of events (counts) occurring in a specific time interval are of interest. Due to missingness, this endpoint is not observed for all patients and the classical approach of analysis will be complete cases which is valid only under MCAR missingness mechanism. However, dropout is usually outcome related and the MCAR assumption becomes untenable.

We proposed the use of two frequentist based imputation methods, asymptotic MLE and bootstrap methods for dealing with missing data in recurrent event data framework. The recurrent events are modeled as over-dispersed Poisson data, with constant rate function. We observed that the use of 10 imputations is sufficient for both methods when the fraction of missing data is up to 50%. The bootstrap approach is recommended in 'very' small samples as the MLEs have large variances and this can reduce precision for asymptotic methods. Although the usual assumption in practice is mixed-Poisson process, numerical instability in the estimation of the variance of the random effects is a pointer to the fact that the assumption may be inappropriate.

8.2 Future Work

The following list gives possible extensions of the work described in this thesis.

1. The SSNM model discussed in chapter 4 inherited its inferential drawbacks from the Azzalini (1985) SN model, therefore, there is need for developments of rigorous inferential tools for the SSNM model. These drawbacks are associated with the skewness parameter, λ in which it can diverge. The most satisfactory method to alleviate this problem so far is the Sartori (2006) modified likelihood. Sartori (2006) used method for bias prevention of the maximum likelihood estimator proposed by Firth (1993). The method modifies the score function such that the resulting estimator has lower bias than the maximum likelihood estimator. The major advantage of this method is its finiteness. Our future work on the SSNM model will use this modified likelihood and also propose Bayesian techniques for parameter estimation.

2. The focus of this thesis has been on modeling skewness. An extension of the SSNM model using skew-t distribution is likely to be a more rewarding exercise since it has the SSNM model as a special case. The challenge in multivariate extensions of the SSNM model is modeling the covariance structure over time. Bounded data requires correlation to decrease with increase in time between measurements, and correlation to increase as the study progress, and outcome attain their final levels. Also, a model for recovery rate and final recovery level is valuable and this can be done when the data is used in longitudinal settings.

3. We did not proceed with the estimation of parameters in the multilevel extension of the SSNM model in section 5.4 because the likelihood function is difficult to evaluate. Our future work will explore the use of Pseudo-likelihood methodology for parameter estimation and develop sensitivity analyses tools for the hypothesis of selection in this settings.

4. The contour plot in Figure 6.2 with SHASH marginals points to the possibility that the Gaussian copula used in chapter 6 may be inappropriate. Although copula functions are theoretically independent of marginals, the geometrical behavior of the marginal densities (being increasing, decreasing, constant or unimodal functions), have influence on the dependence structure. Our future work will search for a new class of dependence functions that will take into account the type of marginals used.

5. It can also be of interest to evaluate the use of multiple imputation in sample selection settings. In principle, there are situations where a variable is skewed and yet the residuals are approximately normal when the skewed variable is conditioned on other variables. In this case, a correctly specified conditional normal imputation can be used. This however, does not remove the effect of boundedness of the scores. Imputation of bounded values changes the mean and variance of the imputed variable. The use of truncated distributions can ensure imputations are done within a specified bounds. Although multiple imputation is valid under MAR assumption, MNAR counterparts can be investigated.

6. There are still open questions on imputation for recurrent event data. For instance, it may be of interest to compare the performance of Bayesian Multiple Imputation to frequentist approach. It could also be of interest to investigate how the use of higher event rates than the MAR affects power in sections 7.4.5 and 7.4.6.

Appendix A

Supplementary Material

A.1 Derivation of Gradients and Observed information matrix

The gradient of the selection skew-normal model log-likelihood can be derived as follows:

$$
\begin{aligned}
\frac{\partial l}{\partial \beta} =& S_i\Bigg(\sum_{i=1}^{n}\Bigg\{\frac{1}{\sigma}z_i - \frac{\lambda}{\sigma}K_3 - \frac{\rho}{\sigma r^{1/2}}K_1\Bigg\}x_i\Bigg)\\
\frac{\partial l}{\partial \gamma} =& S_i\Bigg(\frac{1}{\sqrt{1-\rho^2}r^{1/2}}\sum_{i=1}^{n}K_1x_i\Bigg) + (1-S_i)\Bigg(\sum_{i=1}^{n}(-2)K_2x_i\Bigg)\\
\frac{\partial l}{\partial \sigma} =& S_i\Bigg(\sum_{i=1}^{n}\Bigg\{-\frac{n}{\sigma} + \frac{1}{\sigma}z_i^2 - \frac{\lambda}{\sigma}K_3z_i - \frac{\rho}{\sigma r^{1/2}}K_1z\Bigg\}\Bigg)\\
\frac{\partial l}{\partial \rho} =& S_i\Bigg(\sum_{i=1}^{n}\frac{1}{r^{3/2}}K_1\Big(\rho\gamma' x_i + z_i\Big)\Bigg) + (1-S_i)\Bigg(\sum_{i=1}^{n}\frac{-2\lambda}{\sqrt{2\pi u}}K_4\Bigg)\\
\frac{\partial l}{\partial \lambda} =& S_i\Bigg(\sum_{i=1}^{n}K_3z_i\Bigg) + (1-S_i)\Bigg(\sum_{i=1}^{n}\frac{-2\rho}{(1+\lambda^2)\sqrt{2\pi u}}K_4\Bigg),
\end{aligned}
$$

where, $r = (1-\rho^2)$, $z = (y-\beta' x_i)/\sigma$, $u = (1+\lambda^2-\lambda^2\rho^2)$, and

$$
\begin{cases}
\omega = \frac{\gamma' x_i + \rho\Big(\frac{(y_i-\beta' x_i)}{\sigma}\Big)}{\sqrt{1-\rho^2}} \;\&\; K_1 = \phi(\omega)/\Phi(\omega), \quad K_2 = \frac{\phi\Big(\gamma' x_i\Big)\Phi\Big(\frac{-\gamma' x_i\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\Big)}{\Phi_{SN}\Big(-\gamma' x_i;0,1,\frac{\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\Big)} \\
\zeta = \lambda\Big(\frac{y_i-\beta' x_i}{\sigma}\Big) \;\&\; K_3 = \phi(\zeta)/\Phi(\zeta), \qquad K_4 = \frac{\phi\Big(\frac{\gamma' x_i\sqrt{1+\lambda^2}}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\Big)}{\Phi_{SN}\Big(-\gamma' x_i;0,1,\frac{\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\Big)}
\end{cases}
$$

Note that the derivative of $\Phi_{SN}\left(-\gamma' x_i; 0, 1, \frac{\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\right)$ w.r.t. γ follows the usual differentiation of CDF to get the PDF. However, the derivatives of ρ and γ in this expression is not a straightforward application of this principle. The approach we followed is to re-write the CDF above as a standard bivariate normal integral $\left(2\Phi_2\left(-\gamma' x_i, 0; -\lambda\rho/\sqrt{1+\lambda^2}\right)\right)$. We make use of the fact that, if $\Phi_2(.,.;\rho)$ and $\phi_2(.,.;\rho)$ are standard bivariate normal CDF and PDF respectively, then $\frac{d\Phi_2(.,.;\rho)}{d\rho} = \phi_2(.,.;\rho)$.

The elements of the observed information matrix are:

$$\frac{-\partial^2 l}{\partial\beta^2} = S_i\left(\sum_{i=1}^{n}\left\{\frac{1}{\sigma^2}+\frac{\lambda^2}{\sigma^2}\left[\zeta K_3+K_3^2\right]+\frac{\rho^2}{\sigma^2 r}\left[\omega K_1+K_1^2\right]\right\}x_i^2\right)$$

$$\frac{-\partial^2 l}{\partial\gamma^2} = S_i\left(\sum_{i=1}^{n}\frac{1}{r}\left\{\omega K_1+K_1^2\right\}x_i^2\right)+(1-S_i)\left(\sum_{i=1}^{n}-\left\{2\gamma' x_i K_2-\frac{2\lambda\rho}{\sqrt{2\pi u}}K_4-4K_2^2\right\}x_i^2\right)$$

$$\frac{-\partial^2 l}{\partial\sigma^2} = S_i\left(\sum_{i=1}^{n}\left\{-\frac{n}{\sigma^2}+\frac{3}{\sigma^2}z_i^2-\frac{1}{\sigma^2}\left[2\zeta K_3-\zeta^3 K_3-\zeta^2 K_3^2\right]\right.\right.$$
$$\left.\left.-\frac{\rho}{\sigma^2 r^{1/2}}\left[2z_i K_1-\frac{\rho}{r^{1/2}}z_i^2\omega K_1-\frac{\rho}{r^{1/2}}z_i^2 K_1^2\right]\right\}\right)$$

$$\frac{-\partial^2 l}{\partial\rho^2} = S_i\left(\sum_{i=1}^{n}\left\{-\frac{3\rho}{r^2}\Big(\rho\gamma' x_i+z_i\Big)K_1-\frac{\gamma' x_i}{r}K_1+\frac{1}{r^{5/2}}\Big(\rho\gamma' x_i+z_i\Big)^2\omega K_1+\frac{1}{r^{5/2}}\Big(\rho\gamma' x_i+z_i\Big)^2 K_1^2\right\}\right)$$
$$+(1-S_i)\left(\sum_{i=1}^{n}\left\{\frac{2\lambda^3\rho}{\sqrt{2\pi u^3}}K_4-\frac{2\lambda^3\rho(1+\lambda^2)(\gamma' x_i)^2}{\sqrt{2\pi u^5}}K_4+\frac{4\lambda^2}{2\pi u}K_4^2\right\}\right)$$

$$\frac{-\partial^2 l}{\partial\lambda^2} = S_i\left(\sum_{i=1}^{n}\left\{\zeta K_3+K_3^2\right\}z_i\right)+(1-S_i)\left(\sum_{i=1}^{n}\left\{\frac{-2\lambda\rho^3(\gamma' x_i)^2}{(1+\lambda^2)\sqrt{2\pi u^5}}K_4-\frac{2\lambda\rho(1-\rho^2)}{(1+\lambda^2)\sqrt{2\pi u^3}}K_4\right.\right.$$
$$\left.\left.-\frac{4\lambda\rho}{(1+\lambda^2)^2\sqrt{2\pi u}}K_4+\frac{4\rho^2}{(1+\lambda^2)^2 2\pi u}K_4^2\right\}\right)$$

$$\frac{-\partial^2 l}{\partial\beta\partial\gamma} = S_i\left(\sum_{i=1}^{n}\left\{-\frac{\rho}{\sigma r}\omega K_1-\frac{\rho}{\sigma r}K_1^2\right\}x_i^2\right)$$

$$\frac{-\partial^2 l}{\partial\beta\partial\sigma} = S_i\left(\sum_{i=1}^{n}\left\{\frac{2}{\sigma^2}z_i+\frac{\lambda}{\sigma^2}\zeta^2 K_3-\frac{\lambda}{\sigma^2}K_3+\frac{\lambda}{\sigma^2}\zeta K_3^2+\frac{\rho^2}{\sigma^2 r}z_i\omega K_1-\frac{\rho}{\sigma^2 r^{1/2}}K_1+\frac{\rho^2}{\sigma^2 r}z_i K_1^2\right\}x_i\right)$$

$$\frac{-\partial^2 l}{\partial\beta\partial\rho} = S_i\left(\sum_{i=1}^{n}\left\{-\frac{\rho}{\sigma r^2}\Big(\rho\gamma' x_i+z_i\Big)\omega K_1+\frac{1}{\sigma r^{3/2}}K_1-\frac{\rho}{\sigma r^2}\Big(\rho\gamma' x_i+z_i\Big)K_1^2\right\}x_i\right)$$

$$\frac{-\partial^2 l}{\partial\beta\partial\lambda} = S_i\left(\sum_{i=1}^{n}\left\{\frac{1}{\sigma}K_3-\frac{1}{\sigma}\zeta^2 K_3-\frac{1}{\sigma}\zeta K_3^2\right\}x_i\right)$$

$$\frac{-\partial^2 l}{\partial\gamma\partial\sigma} = S_i\left(\sum_{i=1}^{n}\left\{-\frac{\rho}{\sigma r}\omega K_1 z_i-\frac{\rho}{\sigma r}K_1^2 z_i\right\}x_i\right)$$

$$\frac{-\partial^2 l}{\partial\gamma\partial\rho} = S_i\left(\sum_{i=1}^{n}\left\{\frac{1}{r^2}\Big(\rho\gamma' x_i+z_i\Big)\omega K_1-\frac{\rho}{r^{3/2}}K_1+\frac{1}{r^2}\Big(\rho\gamma' x_i+z_i\Big)K_1^2\right\}x_i\right)$$
$$+(1-S_i)\left(\sum_{i=1}^{n}\left\{\frac{4\lambda}{\sqrt{2\pi u}}\phi\left(\frac{\gamma' x_i\sqrt{1+\lambda^2}}{u^{1/2}}\right)K_2-\frac{2\lambda(1+\lambda^2)}{\sqrt{2\pi u^3}}(\gamma' x_i)K_4\right\}x_i\right)$$

$$\frac{-\partial^2 l}{\partial\gamma\partial\lambda} = (1-S_i)\left(\sum_{i=1}^{n}\left\{\frac{4\rho}{(1+\lambda^2)\sqrt{2\pi u}}\phi\left(\frac{\gamma' x_i\sqrt{1+\lambda^2}}{u^{1/2}}\right)K_2-\frac{2\rho}{\sqrt{2\pi u^3}}(\gamma' x_i)K_4\right\}x_i\right)$$

$$\frac{-\partial^2 l}{\partial\sigma\partial\rho} = S_i\left(\sum_{i=1}^{n}\left\{\frac{1}{\sigma r^{3/2}}K_1 z_i-\frac{\rho}{\sigma r^2}\Big(\rho\gamma' x_i+z_i\Big)\omega K_1 z_i-\frac{\rho}{\sigma r^2}\Big(\rho\gamma' x_i+z_i\Big)K_1^2 z_i\right\}\right)$$

$$\frac{-\partial^2 l}{\partial\sigma\partial\lambda} = S_i\left(\sum_{i=1}^{n}\left\{\frac{1}{\sigma}K_3 z_i - \frac{1}{\sigma}\zeta^2 K_3 z_i - \frac{1}{\sigma}\zeta K_3^2 z_i\right\}\right)$$

$$\frac{-\partial^2 l}{\partial\rho\partial\lambda} = (1-S_i)\left(\sum_{i=1}^{n}\left\{\frac{2}{\sqrt{2\pi u^3}}K_4 - \frac{2\lambda^2\rho^2}{\sqrt{2\pi u^5}}(\gamma' x_i)^2 K_4 + \frac{4\lambda\rho}{2\pi u(1+\lambda^2)}K_4^2\right\}\right).$$

A.2 Simulation results for fixed λ and varying ρ

In this section, the effects of varying correlation coefficient ρ when λ is fixed to be 1 and 2 is considered. The results in Tables A.1 and A.2 show that the bias in the estimate of λ decreases as the strength of the correlation increases. This is in line with the fact that both ρ and λ contribute to the skewness present in the observed data.

Table A.1: Simulation results (multiplied by 10,000) for $\lambda = 1$ and varying ρ in the presence of exclusion restriction.

		Bias			MSE		
		SSNM	SNM	TS	SSNM	SNM	TS
$\rho = 0.0$	β_0	990	5636	5637	977	3196	3197
	β_1	7	4	3	14	14	14
	γ_0	-35	72	73	110	54	54
	γ_1	49	75	76	63	63	63
	γ_2	109	148	149	101	102	102
	σ	-385	-1746	-1746	121	310	310
	ρ	42	22	15	194	145	143
	λ	-2164	-	-	3274	-	-
$\rho = 0.3$	β_0	461	5627	5630	384	3183	3187
	β_1	7	9	7	13	13	13
	γ_0	209	1960	1965	181	448	451
	γ_1	58	226	233	68	72	73
	γ_2	114	366	376	111	119	121
	σ	-113	-1714	-1714	71	299	299
	ρ	-26	-432	-449	173	153	158
	λ	-541	-	-	1529	-	-
$\rho = 0.7$	β_0	484	5614	5619	375	3164	3171
	β_1	1	9	6	11	11	12
	γ_0	637	5395	5437	478	3011	3065
	γ_1	185	1036	1078	93	187	202
	γ_2	309	1583	1645	173	383	412
	σ	-123	-1684	-1683	67	289	289
	ρ	-93	-656	-683	70	113	165
	λ	-564	-	-	1518	-	-

Table A.2: Simulation results (multiplied by 10,000) for $\lambda = 2$ and varying ρ in the presence of exclusion restriction.

		Bias			MSE		
		SSNM	SNM	TS	SSNM	SNM	TS
$\rho = 0.0$	β_0	9	7127	7130	41	5093	5097
	β_1	10	4	4	9	30	30
	γ_0	-4	72	73	193	54	54
	γ_1	6	74	76	62	63	63
	γ_2	45	148	149	100	102	102
	σ	-4	-2996	-2996	27	902	902
	ρ	13	30	13	265	138	135
	λ	339	-	-	1110	-	-
$\rho = 0.3$	β_0	6	7107	7114	38	5063	5073
	β_1	10	13	30	8	9	9
	γ_0	40	2538	2543	256	714	717
	γ_1	22	325	333	73	80	81
	γ_2	57	509	524	125	138	140
	σ	-1	-2918	-2919	23	856	856
	ρ	-14	-739	-776	224	200	203
	λ	331	-	-	1034	-	-
$\rho = 0.7$	β_0	24	7072	7079	34	5010	5021
	β_1	5	26	21	7	8	8
	γ_0	251	7605	7683	311	5911	6038
	γ_1	141	1754	1832	102	400	432
	γ_2	226	2629	2757	198	849	922
	σ	-12	-2851	-2850	21	817	816
	ρ	-86	-1162	-1194	88	236	295
	λ	284	-	-	991	-	-

A.3 PDFs and h-functions of some selected copulas

Bivariate t-copula

$$c(u_1, u_2; \rho, \eta) = \frac{\Gamma\left(\frac{\eta+2}{2}\right)\Gamma\left(\frac{\eta}{2}\right)}{\sqrt{1-\rho^2}\left[\Gamma\left(\frac{\eta+2}{2}\right)\right]^2}\left(1+\frac{x_1^2}{\eta}\right)^{\frac{\eta+1}{2}}\left(1+\frac{x_2^2}{\eta}\right)^{\frac{\eta+1}{2}}\left(1+\frac{x_1^2+x_2^2-2\rho x_1 x_2}{\eta(1-\rho^2)}\right)^{-\frac{\eta+2}{2}}$$

$$h(u_1, u_2; \rho, \eta) = t_{\eta+1}\left[(x_1-\rho x_2)\left(\frac{\eta+x_2^2(1-\rho^2)}{\eta+1}\right)^{-1/2}\right], \text{where } x_1 = t_\eta^{-1}(u_1) \text{ and } x_2 = t_\eta^{-1}(u_2).$$

Bivariate Clayton copula

$$c(u_1,u_2;\delta) = (1+\delta)(u_1u_2)^{-(1+\delta)}\Big(u_1^{-\delta}+u_2^{-\delta}-1\Big)^{-1/\delta-2}$$
$$h(u_1,u_2;\delta) = u_2^{-(1+\delta)}\Big(u_1^{-\delta}+u_2^{-\delta}-1\Big)^{-(1+1/\delta)}.$$

Bivariate Gumbel copula

$$\begin{aligned} c(u_1,u_2;\delta) =& C(u_1,u_2;\delta)(u_1u_2)^{-1}\Big[(-\log u_1)^\delta+(-\log u_2)^\delta\Big]^{-2+2/\delta}(\log u_1\log u_2)^{\delta-1} \\ &\times\Big[1+(\delta-1)\Big((-\log u_1)^\delta+(-\log u_2)^\delta\Big)^{-1/\delta}\Big] \\ h(u_1,u_2;\delta) =& C(u_1,u_2;\delta).\frac{1}{u_2}.\Big(-\log u_2\Big)^{\delta-1}\Big((-\log u_1)^\delta+(-\log u_2)^\delta\Big)^{1/\delta-1}. \end{aligned}$$

A.4 R-codes for copula based truncated sample selection model

```
shashlike <- function(bstart,y1,x1,y2,x2,a,b){
if (match("gamlss",.packages(),0)==0) require(gamlss)

dtrunc <- function(x, spec, a = a, b = b, ...)# defining general truncated PDFs
{
tt <- rep(0, length(x))
g <- get(paste("d", spec, sep = ""), mode = "function")
G <- get(paste("p", spec, sep = ""), mode = "function")
tt[x>=a & x<=b] <- g(x[x>=a&x<=b], ...)/(G(b, ...) - G(a, ...))
return(tt)
}

ptrunc <- function(x, spec, a = a, b = b, ...)# defining general truncated CDFs
{
tt <- x
aa <- rep(a, length(x))
bb <- rep(b, length(x))
G <- get(paste("p", spec, sep = ""), mode = "function")
tt <- G(apply(cbind(apply(cbind(x, bb), 1, min), aa), 1, max), ...)
tt <- tt - G(aa, ...)
tt <- tt/(G(bb, ...) - G(aa, ...))
```

```
return(tt)
}
p=ncol(x1); k=ncol(x2)
b1 =bstart[1:p];b2 =bstart[(p+1):(k+p)]
sigma <- bstart[(k+p+1)]
if (sigma < 0)
    return(NA)
rho <- bstart[k+p+2]# Linear correlation from Gaussian copula
nu <- bstart[k+p+3]# skenwess parameter of SHASH model
if ((rho < -1) || (rho > 1))
    return(NA)
xb1 = x1%*%b1; xb2 = x2%*%b2; u2 <- y2-xb2; r <- sqrt(1 - rho^2)
Bb1 <- qnorm(pnorm(xb1))
Bb2 <-rho*qnorm(ptrunc(y2,"SHASHo", a = a, b = b,mu =xb2, sigma = sigma,
nu= nu, tau= 1))
Bb <- Bb1+Bb2; B <- Bb/r
l1 <- dtrunc(y2,"SHASHo", a = a, b = b,mu = xb2, sigma = sigma, nu= nu, tau = 1)
l1 <- log(l1)
b <- log(1-pnorm(xb1))
ll<- ifelse(y1==0,b,l1+(pnorm(B,log.p=TRUE)))
return(-sum(ll))
}
```

A.5 Tables for Part II of the thesis

Table A.3: Imputation with new rate $\lambda_{\text{new,trt}}(t)$. 30% data is missing in both the treated and the placebo arm.

			Parameters			Std. Err.		
Size	Rate		ϕ	δ	β	ϕ	δ	β
		True Val.	0.5000	0.0200	-0.3000	-	-	-
n=400	$\lambda = 1.05$	Asymptotic	0.5138	0.0202	-0.2793	0.1754	0.0693	0.1117
		Bootstrap	0.5033	0.0201	-0.2793	0.1639	0.0689	0.1016
	$\lambda = 1.10$	Asymptotic	0.5279	0.0202	-0.2596	0.1737	0.0698	0.1125
		Bootstrap	0.5167	0.0201	-0.2598	0.1612	0.0694	0.1020
	$\lambda = 1.20$	Asymptotic	0.5569	0.0202	-0.2201	0.1705	0.0708	0.1143
		Bootstrap	0.5445	0.0201	-0.2219	0.1559	0.0704	0.1029
	$\lambda = 1.50$	Asymptotic	0.6542	0.0202	-0.0984	0.1650	0.0741	0.1222
		Bootstrap	0.6392	0.0201	-0.1021	0.1411	0.0737	0.1060
n=1000	$\lambda = 1.05$	Asymptotic	0.5129	0.0201	-0.2834	0.1083	0.0438	0.0701
		Bootstrap	0.5094	0.0201	-0.2829	0.1010	0.0437	0.0644
	$\lambda = 1.10$	Asymptotic	0.5268	0.0201	-0.2639	0.1072	0.0441	0.0706
		Bootstrap	0.5230	0.0201	-0.2635	0.0993	0.0440	0.0647
	$\lambda = 1.20$	Asymptotic	0.5557	0.0201	-0.2245	0.1051	0.0447	0.0718
		Bootstrap	0.5504	0.0201	-0.2250	0.0962	0.0446	0.0652
	$\lambda = 1.50$	Asymptotic	0.6527	0.0201	-0.1042	0.1022	0.0469	0.0766
		Bootstrap	0.6472	0.0201	-0.1059	0.0873	0.0467	0.0673

Table A.4: Imputation with new rate $\lambda_{\text{new,trt}}(t)$, 10% and 30% data is missing in placebo and treated arm respectively.

			Parameters			Std. Err.		
Size	Rate		ϕ	δ	β	ϕ	δ	β
		True Val.	0.5000	0.0200	-0.3000	-	-	-
n=400	$\lambda = 1.05$	Asymptotic	0.5104	0.0201	-0.2780	0.1671	0.0691	0.1088
		Bootstrap	0.5034	0.0200	-0.2805	0.1624	0.0689	0.1015
	$\lambda = 1.10$	Asymptotic	0.5243	0.0201	-0.2582	0.1652	0.0696	0.1101
		Bootstrap	0.5163	0.0200	-0.2620	0.1598	0.0694	0.1018
	$\lambda = 1.20$	Asymptotic	0.5539	0.0201	-0.2183	0.1615	0.0707	0.1124
		Bootstrap	0.5459	0.0200	-0.2215	0.1542	0.0704	0.1028
	$\lambda = 1.50$	Asymptotic	0.6524	0.0201	-0.0963	0.1544	0.0741	0.1212
		Bootstrap	0.6422	0.0200	-0.1017	0.1396	0.0738	0.1060
n=1000	$\lambda = 1.05$	Asymptotic	0.5102	0.0200	-0.2795	0.1039	0.0437	0.0688
		Bootstrap	0.5077	0.0200	-0.2803	0.1007	0.0437	0.0643
	$\lambda = 1.10$	Asymptotic	0.5243	0.0200	-0.2597	0.1026	0.0441	0.0694
		Bootstrap	0.5215	0.0200	-0.2603	0.0991	0.0440	0.0645
	$\lambda = 1.20$	Asymptotic	0.5542	0.0200	-0.2195	0.1005	0.0447	0.0709
		Bootstrap	0.5508	0.0200	-0.2209	0.0958	0.0447	0.0651
	$\lambda = 1.50$	Asymptotic	0.6508	0.0200	-0.0995	0.0961	0.0468	0.0762
		Bootstrap	0.6465	0.0200	-0.1019	0.0870	0.0467	0.0672

Table A.5: Imputation with new rate $\lambda_{\text{new,trt}}(t)$, 10% and 40% data is missing in placebo and treated arms respectively.

			Parameters			Std. Err.		
Size	Rate		ϕ	δ	β	ϕ	δ	β
		True Val.	0.5000	0.0200	-0.3000	-	-	-
n=400	$\lambda = 1.05$	Asymptotic	0.5142	0.0201	-0.2705	0.1688	0.0693	0.1129
		Bootstrap	0.5073	0.0200	-0.2732	0.1619	0.0690	0.1017
	$\lambda = 1.10$	Asymptotic	0.5308	0.0201	-0.2454	0.1667	0.0699	0.1146
		Bootstrap	0.5230	0.0200	-0.2487	0.1587	0.0696	0.1020
	$\lambda = 1.20$	Asymptotic	0.5647	0.0201	-0.1960	0.1630	0.0711	0.1177
		Bootstrap	0.5560	0.0200	-0.1998	0.1526	0.0708	0.1031
	$\lambda = 1.50$	Asymptotic	0.6729	0.0201	-0.0499	0.1559	0.0747	0.1278
		Bootstrap	0.6619	0.0200	-0.0557	0.1373	0.0744	0.1065
n=1000	$\lambda = 1.05$	Asymptotic	0.5126	0.0200	-0.2733	0.1048	0.0438	0.0712
		Bootstrap	0.5094	0.0200	-0.2746	0.1006	0.0437	0.0643
	$\lambda = 1.10$	Asymptotic	0.5290	0.0200	-0.2484	0.1035	0.0442	0.0720
		Bootstrap	0.5257	0.0200	-0.2498	0.0986	0.0441	0.0646
	$\lambda = 1.20$	Asymptotic	0.5621	0.0200	-0.1998	0.1015	0.0449	0.0741
		Bootstrap	0.5601	0.0200	-0.1999	0.1001	0.0449	0.0685
	$\lambda = 1.50$	Asymptotic	0.6701	0.0200	-0.0542	0.0971	0.0472	0.0805
		Bootstrap	0.6657	0.0200	-0.0567	0.0855	0.0472	0.0674

Bibliography

Aas, K. (2005). *Modelling the dependence structure of financial assets: A survey of four copulas*. Last accessed January 07, 2013 at `http://www.nr.no/files/samba/bff/SAMBA2204c.pdf`.

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics* **44**, 182198.

Ahn, S. C. (1992). The lagrangean multiplier test for a model with two selectivity criteria. *Economics Letters* **38**, 9–15.

Akacha, M. and N. Benda (2010). The impact of dropouts on the analysis of dose-finding studies with recurrent event data. *Statistics in Medicine* **29**, 1635–1646.

Andersen, P. K. and R. Gill (1982). Cox's regression model for counting process: a large sample study. *The Annals of Statistics* **10**, 1100–1120.

Andrews, D. F. and A. M. Herzberg (1985). *A Collection of Problems from Many Fields for the Student and Research Worker*. New York: Springer-Verlag.

Arellano-Valle, R. B. and A. Azzalini (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* **33**, 561–574.

Arellano-Valle, R. B., M. D. Branco, and M. G. Genton (2006). A unified view of skewed distributions arising from selections. *The Canadian Journal of Statistics* **34**, 581–601.

Arellano-Valle, R. B., G. del Pino, and E. San Martin (2002). Definition and probabilistic properties of skew-distributions. *Statistics and Probability Letters* **58**, 111–121.

Arellano-Valle, R. B. and M. G. Genton (2010). Multivariate extended skew-t distributions and related families. *METRON - International Journal of Statistics* **68**, 201–234.

Arellano-Valle, R. B., H. W. Gomez, and F. A. Quintana (2004). A new class of skew-normal distributions. *Communications in Statistics: Theory and Methods* **33**, 1465–1480.

Arendt, J. N. and A. Holm (2006). Probit models with binary endogenous regressors. *Department of Business and Economics, University of Southern Denmark. Paper No.* $\mathbf{4-2006}$, 1.

Arnold, B. C. and R. J. Beaver (2000). Hidden truncation models. *Sankhya Series A* **62**, 23–35.

Arnold, B. C. and R. J. Beaver (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test* **11**, 7–54.

Arnold, B. C. and R. J. Beaver (2007). Skewing Around: Relationships among Classes of Skewed Distributions. *Methodology and Computing in Applied Probability* **9**, 153–162.

Arnold, B. C., R. J. Beaver, R. A. Groeneveld, and W. Q. Meeker (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika* **58**, 471–488.

Arnold, B. C. and R. A. Groeneveld (1995). Measuring skewness with respect to the mode. *The American Statistician* **49**, 34–38.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171–178.

Azzalini, A. and A. Dalla Valle (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–726.

Balakrishnan, N. and X. Zhao (2009). New multi-sample nonparametric tests for panel count data. *The Annals of Statistics* **37**, 1112–1149.

Barnard, J. and D. B. Rubin (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86**, 948–955.

Bellio, R. and E. Gori (2003). Impact evaluation of job training programmes: Selection bias in multilevel models. *Journal of Applied Statistics* $\mathbf{30:8}$, 893–907.

Birnbaum, Z. W. (1950). Effect of linear truncation on a multinormal population. *The Annals of Mathematical Statistics* **21**, 272–279.

Burton, K., T. McClune, and G. Waddell (2001). *The Whiplash Book*. London: TSO, The Stationery Office.

Capitanio, A., A. Azzalini, and E. Stanghellini (2003). Graphical models for skew-normal variates. *Scandinavian Journal of Statistics* **30**, 129–144.

Carpenter, J., S. Pocock, and C. J. Lamm (2002). Coping with missing data in clinical trials: A model-based approach applied to asthma trials. *Statistics in Medicine* **21**, 1043–1066.

Cartinhour, J. (1990). One-dimensional marginal density functions of a truncated multivariate normal density function. *Communications in Statistics- Theory and Methods* **19**, 197–203.

Collins, L., J. L. Schafer, and C. Kam (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* **6**, 330–351.

Cook, R. J. and J. F. Lawless (2002). Analysis of repeated events. *Statistical Methods in Medical Research* **11**, 141–166.

Cook, R. J. and J. F. Lawless (2007). *The Statistical Analysis of Recurrent Events*. Berlin: Springer.

Copas, J. B. and H. Li (1997). Inference for non-random samples. *Journal of the Royal Statistical Society: Series B* **59**, 55–95.

Cox, D. R. (1972). Regression models and life-tables. *Journal of Royal Statistical Society, Series B* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Dean, C. B. and R. Balshaw (1997). Efficiency lost by analyzing counts rather than event times in poisson and overdispersed poisson regression models. *Journal of the American Statistical Association* **92**, 1387–1398.

Diggle, P. and M. G. Kenward (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics* **43**, 49–93.

Dominguez-Molina, J. A., G. Gonzalez-Farias, and R. Ramos-Quiroga (2004). Skew-normality in stochastic frontier analysis. In M. G. Genton (Ed.), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, pp. 223–241. Boca Raton, Florida: Chapman & Hall, CRC.

Escarela, G. and J. F. Carriere (2003). Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research* **12**, 333–349.

Fernandez, C. and M. F. J. Steel (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* **93**, 359–371.

Ferreira, J. T. A. S. and M. F. J. Steel (2007). A new class of skewed multivariate distributions with applications to regression analysis. *Statistica Sinica* **17**, 505–529.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Flecher, C., D. Allard, and P. Naveau (2010). Truncated skew-normal distributions: moments, estimation by weighted moments and application to climate data. *METRON- International Journal of Statistics* **68**, 331–345.

Genest, C. and A. C. Favre (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* **12**, 347–368.

Genest, C., K. Ghoudi, and L. -P. Rivest (1995). A semiparametric estimation procedure of dependent parameters in multivariate families of distributions. *Biometrika* **82**, 543–552.

Genius, M. and E. Strazzera (2004). The copula approach to sample selection modelling: an application to the recreational value of forests. *FEEM Working Paper. Paper No.* $\mathbf{73 - 04}$.

Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality.* Boca Raton, Florida: Chapman & Hall, CRC.

Genton, M. G., Y. Ma, and H. Sang (2011). On the likelihood function of gaussian max-stable processes. *Biometrika* **98**, 481–488.

Gonzalez-Farias, G., J. A. Dominguez-Molina, and A. K. Gupta (2004). The closed skew-normal. In M. G. Genton (Ed.), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, pp. 25–42. Boca Raton, Florida: Chapman & Hall, CRC.

Graham, J. W., A. E. Olchowski, and T. D. Gilreath (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* **8**, 206–213.

Gupta, A. K., F. C. Chang, and W. J. Huang (2002). Some skew-symmetric models. *Random Operators and Stochastic Equations* **10**, 133–140.

Gupta, A. K., Graciela Gonzalez-Farias, and J. Armando Dominguez-Molina (2004). A multivariate skew normal distribution. *Journal of Multivariate Analysis* **89**, 181–190.

Hallin, M. and C. Ley (2012). Skew-symmetric distributions and Fisher information a tale of two densities. *Bernoulli* **18**, 747–763.

Ham, J. C. (1982). Estimation of a labour supply model with censoring due to unemployment and underemployment. *Review of Economic Studies* **49**, 335–354.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* **5**, 475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–161.

Henze, N. (1986). A probabilistic representation of the 'skew-normal' distribution. *Scandinavian Journal of Statistics* **13**, 271–275.

Hutton, J. L. and E. Stanghellini (2011). Modelling bounded health scores with censored skew-normal distributions. *Statistics in Medicine* **30**, 368–376.

Jahn-Eimermacher, A. (2008). Comparison of the Andersen-Gill model with Poisson and negative binomial regression on recurrent event data. *Computational Statistics and Data Analysis* **52**, 4989–4997.

Jamalizadeh, A. and N. Balakrishnan (2009). Order statistics from trivariate normal and t_v distributions in terms of generalized skew-normal and skew-t_v distributions. *Journal of Statistical Planning and Inference* **139**, 3799–3819.

Jamalizadeh, A. and N. Balakrishnan (2010). Distributions of order statistics and linear combinations of order statistics from an elliptical distribution as mixtures of unied skew-elliptical distributions. *Journal of Multivariate Analysis* **101**, 1412–1427.

Jamalizadeh, A., J. Behboodian, and N. Balakrishnan (2008). A two-parameter generalized skew-normal distribution. *Statistics and Probability Letters* **78**, 1722–1728.

Jamalizadeh, A., R. Pourmousa, and N. Balakrishnan (2009). Truncated and limited skew-normal and skew-t distributions: Properties and an illustration. *Communications in Statistics - Theory and Methods* **38**, 2653–2668.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Johnson, R. A. and D. W. Wichern (2007). *Applied Multivariate Statistical Analysis* (6 ed.). New Jersey: Prentice Hall, Inc.

Jones, M. C. (2004). Families of distributions arising from distributions of order statistics. *Test* **13**, 1–43.

Jones, M. C. and A. Pewsey (2009). Sinh-arcsinh distributions. *Biometrika* **96**, 761–780.

Kaarik, E. and M. Kaarik (2009). Modeling dropouts by conditional distribution, a copula-based approach. *Journal of Statistical Planning and Inference* **139**, 3830–3835.

Kim, H. (2002). Binary regression with a class of skewed t link models. *Communication in Statistics- Theory and Methods* **31**, 1863–1886.

Kim, H. (2004). A family of truncated skew-normal distributions. *The Korean Communications* **11**, 265–274.

Kotz, S., N. Balakrishnan, and N. L. Johnson (2000). *Continuous Multivariate Distributions- Vol. 1* (2 ed.). New York: John Wiley & Sons Ltd.

Lamb, S. E., S. Gates, M. R. Underwood, M. W. Cooke, D. Ashby, A. Szczepura, M. A. Williams, E. M. Williamson, E. J. Withers, S. M. Isa, and A. Gumber (2007). Managing Injuries of the Neck Trial (MINT): design of a randomised controlled trial of treatments for whiplash associated disorders. *BMC Muscloskeletal Disorder* **8**, :7.

Lambert, P. and F. Vandenhende (2002). A copula-based model for multivariate non-normal londitudinal data: Analysis of a dose titration safety study on a new antidepresant. *Statistics in Medicine* **21**, 3197–3217.

Lawless, J. F. (1987a). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics* **15**, 209–225.

Lawless, J. F. (1987b). Regression methods for poisson process data. *Journal of the American Statistical Association* **82**, 808–815.

Lee, L. (1983). Generalized econometric models with selectivity. *Econometrica* **51**, 507–5012.

Ley, C. and D. Paindaveine (2010). On the singularity of multivariate skew-symmetric models. *Journal of Multivariate Analysis* **101**, 1434–1444.

Lin, D., L. Wei, I. Yang, and Z. Ying (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B* **62**, 711–730.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association* **90**, 1112–1121.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2 ed.). New Jersey: John Wiley & Sons Ltd.

Little, R. J. A. and L. Yau (1996). Intention to treat analysis for longitudinal studies with drop outs. *Biometrics* **52**, 1324–1333.

Loperfido, N. (2002). Statistical implications of selectively reported inferential results. *Statistics and Probability Letters* **56**, 13–22.

Luca, G. D. and F. Peracchi (2006). *A sample selection model for unit and item nonresponse in cross-sectional surveys.* Last accessed July 20, 2010 at `http://www.share-project.org/t3/share/uploads/tx_sharepublications/deLuca_Peracchi_06.pdf`.

Marchenko, Y. V. and M. G. Genton (2012). A Heckman selection-t model. *Journal of the American Statistical Association* **107** : **497**, 304–317.

Meng, C. and P. Schmidt (1985). On the cost of partial observability in the bivariate probit model. *International Economic Review* **26**, 71–85.

Metcalfe, C. and S. G. Thompson (2006). The importance of varying the event generation process in simulation studies of statistical methods for recurrent events. *Statistics in Medicine* **25**, 165–179.

Metcalfe, C. and S. G. Thompson (2007). Wei, Lin and Weissfeld's marginal analysis of multivariate failure data: should it be applied to a recurrent events outcome? *Statistical Methods in Medical Research* **16**, 103–122.

Nadarajah, S. and S. Kotz (2003). Skewed distributions generated by the normal kernel. *Statistics and Probability Letters* **65**, 269–277.

Nadarajah, S. and S. Kotz (2006). R programs for computing truncated distributions. *Journal of Statistical Software* **16**.

Nelsen, R. B. (2006). *An Introduction to Copulas* (2 ed.). New York: Springer.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review* **71**, 593–627.

O'Hagan, A. and T. Leonard (1976). Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* **63**, 201–203.

Pewsey, A. (2000). Problems of inference for Azzalini's skew-normal distribution. *Journal of Applied Statistics* **27**, 859–870.

Pewsey, A. (2006). Some observations on a simple means of generating skew distributions. In B. C. Arnold, N. Balakrishnan, E. Castillo, and J. M. Sarabia (Eds.), *Advances in Distribution Theory, Order Statistics and Inference*, pp. 75–84. Boston: Birkhauser.

Poirier, D. J. (1980). Partial observability in bivariate probit models. *Journal of Econometrics* **12**, 209–217.

Prieger, J. E. (2002). A flexible parametric selection model for non-normal data with application to health care usage. *Journal of Applied Econometrics* **17**, 367–392.

Puhani, P. A. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys* **14**, 53–68.

Roberts, C. (1966). A correlation model useful in the study of twins. *Journal of the American Statistical Association* **61**, 1184–1190.

Robins, J. M. and R. D. Gill (1997). Non-response models for the analysis of nonmonotone ignorable missing data. *Statistics in Medicine* **16**, 39–56.

Robins, J. M. and N. Wang (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.

Rosco, J. F., M. C. Jones, and A. Pewsey (2011). Skew-t distributions via the sinh-arcsinh transformation. *Test* **20**, 630–652.

Rosenman, R., B. Mandal, V. Tennekoon, and L. G. Hill (2010). *Estimating treatment effectiveness with sample selection.* Last accessed December 31, 2011 at `http://faculty.ses.wsu.edu/WorkingPapers/Rosenman/WP2010-5.pdf`.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Survey.* New York: John Wiley & Sons Ltd.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* **57**, 3–18.

Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew-normal and skew-t distributions. *Journal of Statistical Planning and Inference* **136**, 4259–4275.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* New York: Chapman & Hall, CRC.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research* **8**, 3–15.

Schafer, J. L. and J. W. Graham (2002). Missing data: Our view of the State of the Art. *Psychological Methods* **7**, 147–177.

Smith, M. D. (2003). Modelling sample selection using archimedean copulas. *Econometrics Journal* **6**, 99–123.

Sun, J. and L. J. Wei (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society: Series B* **62**, 293–302.

Sun, J. and L. J. Wei (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika* **89**, 39–48.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

Van Burren, S. (2012). *Flexible imputation of missing data.* Boca Raton: Chapman & Hall, CRC.

Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag, Inc.

Vernon, H. (2009). *The Neck Disability Index: An instrument for measuring self-rated disability due to neck pain or whiplash-assocaited disorder.* Last accessed Febuary 20, 2010 at `http://www.cmcc.ca/Portals/0/PDFs/Research_05_2009_NDI_Manual.pdf`.

Vernon, H. and S. Mior (1991). The Neck Disability Index: a study of reliabity and validity. *Journal of Manipulative and Physiological Therapeutics* **7**, 409–415.

Wang, N. and J. M. Robins (1998). Large-sample theory for parametric imputation procedures. *Biometrika* **85**, 935–948.

Wang, W. and M. T. Wells (2000). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association* **95**, 62–72.

Wei, L., D. Lin, and L. Weissfeld (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association* **84**, 1065–1073.

Weinstein, M. A. (1964). The sum of values from a normal and a truncated normal distribution. *Technometrics* **6**, 104–105.

Wellner, J. A. and Y. Zhang (2000). Two estimators of the mean of a counting process with panel count data. *The Annals of Statistics* **28**, 779–814.

Wu, M. C. and K. Bailey (1989). Estimation and comparison of changes in the presence of informative censoring: conditional linear model. *Biometrics* **45**, 938–955.

Wu, M. C. and R. J. Carroll (1988). Estimation and comparison of changes in the presence of informative censoring by modeling the censoring process. *Biometrics* **44**, 175–188.

Zhang, L. and V. P. Singh (2006). Bivariate flood frequency analysis using the copula model. *Journal of Hydrologic Engineering* **11**, 150–164.

Zhang, Y. (2006). Nonparametric k-sample tests for panel count data. *Biometrika* **93**, 777–790.